56èmes Journées de Statistique de la SFdS

du 2 au 6 juin 2025

Recueil des résumés

56èmes Journées de Statistique de la SFdS

à Marseille du 2 au 6 juin 2025

Conférenciers invités :
Julyan Arbel (Inria, Grenoble)
Alice Cleynen (CNRS, Montpellier)
Carlo Gaetan (U Ca Foscari Venise)
Toby Dylan Hocking (U Sherbrooke)
Estelle Kuhn (U Paris Saclay INRAE)
Claire Lacour (U G Eiffel Marne-la-Vallée)
Pierre Alexandre Mattei (Inria, Sophia Antipolis)
Anne Sabourin (U Paris Cité)
Julien Trufin (ULB Bruxelles)
Simon Wood (U Edinburgh)

Conférence Lucien Le Cam :
Judith Rousseau (Oxford & U Paris Dauphine)
Prix Pierre Simon de Laplace :
Sara van de Geer (ETH Zürich)

Présidente du comité scientifique :
Gwladys Toulemonde (U Montpellier, Inria)

Présidente du comité d'organisation :
Mitra Fouladirad (Centrale Méditerrannée)

Aix-Marseille Université | Campus Saint-Charles
https://jds2025.sciencesconf.org
jds2025@sciencesconf.org

Table des matières

Conférence Lucien Le Cam **15**
Bayesian semi-parametric : a curse or a blessing, Rousseau Judith 15

Prix Laplace **17**
Small noise statistics, Van De Geer Sara 17

Prix du Docteur Norbert MARX **19**
Nonlinear multilevel joint model for individual lesion kinetics and survival to characterise intra-individual heterogeneity in patients with advanced cancer, Kerioui Marion 19

Sessions prix étudiants **21**
Modélisation et prédiction du Best Estimate par réseau de neurones génératifs, Ebrotie Kouassi Alphonse Nathan 21
Bien-être des porcs en élevages alternatifs en France, Toti Vito *et al.* 23
Prix SD/SFdS : Mise en place d'une base de données pour la gestion des approvisionnements en matières premières de LOEUF, Durand Anaëlle 24

Sessions invitées semi-plénières **25**
Bayesian deep learning, overview and challenges, Arbel Julyan 25
Processus de Markov déterministes par morceaux et Processus Décisionnels Markoviens, comment les combiner pour mieux soigner., Cleynen Alice 27
Spatio-temporal models that combines different type of extremal dependencies, Gaetan Carlo 28
Algorithmes efficaces pour la détection optimale des ruptures dans les données séquentielles, Hocking Toby 30
Identification et caractérisation de la variabilité individuelle au sein d'une population à partir d'un modèle mécaniste et d'effets mixtes. Applications en amélioration variétale, Kuhn Estelle 31
Modèles de mélange nonparamétriques à variables latentes dépendantes, Lacour Claire 33
Méthodes d'ensembles en apprentissage, Mattei Pierre-Alexandre 34
Statistical learning viewpoints on extreme value analysis, Sabourin Anne 35
Two problems in smooth regression: shape constraint and auto-correlation, Simon Wood 36
Pricing Insurance Contracts and Managing Financial Balances within an Insurance Portfolio, Trufin Julien 37

Hommage à Pierre Cazes **38**
Pierre Cazes : l'éditeur, le statisticien, l'homme, Bry Xavier *et al.* 38

FENStatS **40**
Workshop on FENStatS Accreditation of Statisticians, Pettersson Magnus 40

Session Groupe Jeunes **42**
Notre rapport à la compétition dans l'enseignement supérieur et la recherche : avons-nous besoin de champions ?, Chennetier Guillaume *et al.* 42

Apprentissage & données énergétiques **45**
Spatio-temporal clustering and reconciliation for regional electricity demand forecasting, Brégère Margaux 45
Online learning with signatures for time series forecasting, Drobac Nina 52
Prévision à court terme de consommations électriques locales à l'aide de modèles globaux, Goffinet Roxane 58
Rethinking Synthetic Twins: Average Treatment Effect Estimation with Latent Representations Learning, Ouali Maha 68

Apprentissage fédéré et autres stratégies **77**
Federated Causal Inference: ATE estimation with decentralized observational data, Khellaf Rémi 77
Analyse Raffinée de Federated Averaging et Extrapolation de Richardson-Romberg Fédérée, Mangold Paul 80
Méthodes d'apprentissage faiblement supervisées pour la classification de patients à partir de données de cytométrie en flux, Mikem Pierre-André 81
Leveraging shared information in multimodal generative modeling, Senellart Agathe 88

Apprentissage par méthode bayésienne **94**
Heterogeneous Bayesian Transfer Learning for Complex Ordinary Differential Equations, Abed Youba 94
Réseaux de neurones bayésiens pour la métamodélisation des simulateurs numériques, Baldé Oumar 101
Dirichlet-Based Monte Carlo Dropout for Uncertainty Estimation in Neural Networks, Hoblos Rouaa 107
Régression Logistique Optimale pour la Classification de Distribution Induite par un Arbre, Schall Baptiste 113

Apprentissage par renforcement et plans d'expérience **119**
Apprentissage actif basé sur le gradient des processus gaussiens pour l'analyse de sensibilité globale., Lambert Guerlain 119
Intervalle de confiance asymptotique pour une probabilité estimée à partir d'un plan hypercube latin, Sylvestre Théo 121
Planification d'expériences séquentielle en inférence bayésienne objective, application à l'estimation de courbes de fragilité sismique, Van Biesbroeck Antoine . . . 128
Une approche policy-gradient pour des actions ordinales, Weinberger Simon . . . 134

Apprentissage sur données déséquilibrées **140**
Échantillonnage en deux étapes pour la classification déséquilibrée à l'aide de forêts aléatoires centrées infinies, Mayala Moria Grace Aurore 140
Gestion de données mixtes dans le cadre de données déséquilibrées : application au scoring de clients, Sakho Abdoulaye 146

Data Augmentation with Variational Autoencoder for Imbalanced Dataset, Stocksieker Samuel 155

Biostatistique 1 **163**

Bayesian survival time prediction of French Childhood Cancer Survivors' Study using re-parameterized mixture distribution, Ben Salah Malek 163

Synthèse de données par la méthode Avatar : fidélité en pharmacogénétique, Benoist Clément 170

SurvNMF: Non-Negative Matrix Factorization Supervised for Survival Data Analysis, Le Goff Vincent 176

Challenging estimation of tuning parameters in sPLSDR-Cox models to predict the incidence of rare diseases, Meneghel Danilevicz Ian 182

Biostatistique 2 **188**

Sélection de variables avec contrôle du FDR pour données de grande dimension avec risques concurrents, Cannafarina Hugo 188

Identification supervisée et non supervisée de sous-groupes à l'aide d' un modèle de progression des maladies (Leaspy), Kaisaridi Sofia 196

Testing for misclassified deaths in cancer registries by comparing relative and cause-specific survival, Laverny Oskar 202

Classification **206**

vitesses rapides en classification : au delà du lemme de Zhang, Chardon Hugo . . 206

Set To Be Fair: Set-Valued Classification Under Fairness Constraints, Cohen Eyal 212

Aggrégation d'ensemble de prédiction corrélés par la classe, Fermanian Jean-Baptiste218

Clustering **224**

Sélection du nombre de classes pour un algorithme de clustering à larges marges, Ambroise Christophe *et al.* 224

Classification non supervisée sous contraintes strictes tranchées, Derquenne Christian 231

Segmentation d'images spectrales de fluorescence X sur la base d'une dissimilarité s'appuyant sur la perte de vraisemblance, Megel Maxime 237

Forêts aléatoires non supervisées pour le clustering, Simon Alice 243

Clustering sur graphes et clustering multi-vues **249**

Fusionner plusieurs vues de données par Diffusion Maps, Debaussart Gwendal *et al.* 249

Attributed Graph Partitioning: Comparing K-Means and Optimal Transport in Traffic Applications, Guichard-Sustowski Ketsia 256

The Deep Zero-Inflated Latent Position Block Model for the Clustering of Nodes in Graphs, Niang Seydina Ousmane 262

Méthode intégrée d'inférence de réseaux et de clustering des variables sur des données continues: le modèle Normal-Block, Tous Jeanne 270

Culture statistique (Session groupe Enseignement) **271**

PLUS DE CULTURE STATISTIQUE = PLUS DE CITOYENNETÉ ACTIVE : DONNÉES ET MÉTHODES, Cafieri Simona 271

FUTUR PROGRAMME DE MATHÉMATIQUES DU CYCLE 3 : STATISTIQUE ET PROBABILITÉS EN CM1-CM2-SIXIÈME, Derouet Charlotte 277

50 NUANCES DE STAT : UNE RÉTROSPECTIVE DES CAFÉS DE LA STATISTIQUE À LYON, Rolland Antoine *et al.* 278

Détection d'anomalies **282**

Aggregated Model for Anomaly Detection : A Statistical Learning Approach, Kheirallah Rida *et al.* 282

Density Based Isolation Forest, Leveque Nathan 289

Tester la dynamique maligne des biomarqueurs tumoraux à l'aide des signatures d'un chemin rugueux., Vaucher Rémi 295

Développements théoriques pour l'apprentissage **301**

Sur la non-convexité d'une contrainte de transport et ses conséquences en apprentissage machine, De Lara Lucas 301

General reproducing properties in RKHS with application to derivative operator, El-Boukkouri Fatima-Zahrae 308

On Volume Minimization in Conformal Regression, Le Bars Batiste 313

On Strenghtenings for the Feature Selection Problem, Ma Rui 319

Données de survie, données censurées **326**

Survival Models: Proper Scoring Rule and Stochastic Optimization with Competing Risks, Alberge Julie 326

A general representation of conditional hazard functions in survival analysis, Aurouet Daphné *et al.* 333

Sélection de variables pour données de survie censurées par intervalles, Bercu Ariane339

Apprentissage prédictif en analyse de survie par maximisation empirique de l'indice de concordance de Harrell, Lamalle Florian 345

Données fonctionnelles **350**

Optimal inference for the mean of random functions, Kassi Omar 350

Signal identification by local functional ANOVA, Mahmoud Remi 357

Prédiction du futur de données fonctionnelles à partir de covariables, Sezgin Selman363

Nouvelle approche de scan spatial pour données fonctionnelles : applications économiques et climatiques, Smida Zaineb 371

Régression additive pour réponses Riemaniennes fonctionelles, Van Bever Germain 378

Données longitudinales **383**

Location-scale linear quantile mixed models for fitting longitudinal data with heteroscedasticity, Barbieri Antoine 383

Random Forests using longitudinal predictors, Remiat Justine 389

Time-penalized trees: consistency results and simulations, Valla Mathias *et al.* . 395

Données manquantes **403**

Clustering et réduction de la dimension avec données manquantes, Agliz Yasmine 403

Exploration de la Régression Logistique avec Données Manquantes dans les Covariables, Muller Christophe 410

Une introduction à la classification linéaire avec des données manquantes, Reyero Lobo Angel 418

Données omiques **424**

Inferring cellular heterogeneity with mixture models for DNA methylation rates, Barbot Hugo . . . 424
Sélection de variables sur des données transcriptomiques à l'aide des knockoffs dans un modèle de classification., Cartier Julie . . . 431
Rank-Based Identification of High-dimensional Surrogate Markers: Application to Vaccinology, Hughes Arthur . . . 437
Inférence post hoc pour l'analyse différentielle de données Hi-C, Jorge Elise . . . 446

Données omiques et modèles mixtes **452**
Regularized estimation of non-linear mixed effect models in the presence of high-dimensional biomarkers, Gabaut Auriane . . . 452
phoenics : une nouvelle approche statistique pour l'analyse différentielle longitudinale de voies métaboliques, Guilmineau Camille *et al.* . . . 463
Modèle animal multi-caractères à copule pour améliorer la sélection, Rohmer Tom 469

Enbis & Groupe Fiabilité et incertitudes 1 **475**
Des architectures de réseaux de neurones innovantes pour modéliser la consommation d'électricité, Amara-Ouali Yvenn . . . 475
Etude d'indicateurs de pertinence d'une base de données expérimentales pour la validation de codes de calculs. Application sur la simulation de transitoires thermohydrauliques accidentels, Baccou Jean . . . 484
Prévision de séries temporelles sous contraintes linéaires, Doumèche Nathan . . . 490
Intégration du machine learning dans les processus de prévision de la demande: le cas de Decathlon, Taquet Vianney *et al.* . . . 496

Enbis & Groupe Fiabilité et incertitudes 2 **497**
Modelling and estimation of censored semi-Markov chains, Barbu Vlad . . . 497
MAINTENANCES IMPARFAITES AVEC ACTIONS CORRECTIVES DÉCALÉES : MODÉLISATION, INFÉRENCE STATISTIQUE POUR DONNÉES CENSURÉES ET APPLICATION INDUSTRIELLE, Cousino Theo . . . 504
Optimisation de la vitesse des trains sous contraintes de ponctualité et de consommation énergétique, Funfschilling Christine . . . 510
Étude théorique et empirique de quelques mesures de risque pour les systèmes industriels critiques, Temple-Boyer Marie . . . 514

Environnement et statistique **520**
Un modèle d'EDS contraint pour l'analyse de trajectoires de narvals, Delporte Alexandre . . . 520
Modelling the phytoplankton community in a front: a Mediterranean Sea case study., Garcia Théo . . . 527
Modélisation d'une épidémie végétale à transmission vectorielle sur graphes, application à l'épidémie de Flavescence dorée dans le vignoble de Bourgogne de 2013 à 2023, Roux Jérôme . . . 535
Amélioration de cartes de concentrations en NO2 dans l'air, par combinaison de mesures de qualités différentes, Thulliez Emma . . . 539

Environnement et statistique – Données de pluies **545**
Spatio-temporal generation of precipitations using a spatially correlated Bernoulli and hidden Markov model, Cognot Caroline *et al.* . . . 545

Stochastic rainfall generator with heavy rainfalls and sustained dry spells, Doizé Antoine 553
Distribution composite basée sur le test de Berk-Jones pour la correction de biais saisonnière de précipitations journalières, Ear Philippe 558

Extrêmes et changement climatique **564**
Dynamique des événements extrêmes et processus ponctuels, Heranval Antoine . 564
Exploring Climate Change Through the Lens of Records Theory, Jouni Jinane . . 571
Modélisation de la limite supérieure des températures à la surface de la Terre fondée sur la physique du climat et de leurs variations spatiales, Leterrier Laurie 577

Extrêmes et environnement **584**
Conditions environnementales extrêmes dans le Raz Blanchard, Alves Paul-Adrien 584
Extended extreme-value models with spatial dependence, Dell'oro Lorenzo 591
Simulation of extreme functionals in meteoceanic data, Gorse Nathan 597
Spatio-temporal modeling of urban extreme rainfall events at high resolution, Serre-Combe Chloé 603

Extrêmes et risques **609**
Estimation of extreme risk measures with neural networks, Girard Stéphane . . . 609
CLTs for empirical processes of cluster functionals of extremes in non-mixing time series, Gomez-Garcia José Gregorio 616
Modelling Directional Dependence of Extreme Events, Nicolas Maxime 622
Penalized estimation of GEV parameters for extreme quantile regression, Vidagbandji Lucien M. *et al.* 627

Finance, assurance **633**
Approximate Bayesian Calibration of Option Pricing Models, Blanchard Eléonore 633
Tarification de contrats d'assurance via des données de marchés: Application à l'assurance santé animale, Goffard Pierre-Olivier 641
Combinaison d'assurance traditionnelle et paramétrique : méthode de calibration basée sur l'optimisation d'un critère adapté aux pertes à queue lourde., Nkameni Daniel 648

Graphes et réseaux **651**
Estimation consistante du nombre de clusters dans les modèles à blocs stochastiques multi-couches et dynamiques, Arts Lucie 651
Détection de communauté pour un modèle graphique binaire en grande dimension, Chevallier Julien 658
Génération de modèles graphiques, Gannaz Irène 662
Joint estimation of bipartite network collections. Application to plant-pollinator networks, Lacoste Louis 668

IA générative - Modèles de diffusion **674**
When Are Two Scores Better Than One? Investigating Ensembles of Diffusion Models, Razafindralambo Raphaël 674
Short proof of the optimal convergence of diffusion models, Stéphanovitch Arthur 681
Modèles génératifs basés sur le score et échantillonnage de lois conditionnelles, Strasman Stanislas 683

Taking a Big Step: Large Learning Rates in Denoising Score Matching Prevent Memorization, Wu Yu-Han . . . 689

IA générative - VAE **695**

Clustering guidé dans des autoencodeurs variationels, Courrier Violaine . . . 695

VAE MULTI-COMPOSANT AVEC CHAMP DE MARKOV GAUSSIEN : UNE NOUVELLE APPROCHE POUR LA CONCEPTION INDUSTRIELLE COMPLEXE, Oubari Fouad . . . 702

Analyse Théorique de la Convergence des Autoencodeurs Variationnels, Surendran Sobihan . . . 708

IA pour la découverte scientifique : prospective trans sociétés savantes **715**

AI as a Scientific Pilot?, Alahari Karteek *et al.* . . . 715

AI4Science: From Equations to Algorithms, Gallinari Patrick . . . 718

Les agents conversationnels dans la démarche scientifique : potentiel et limites, Guille Adrien . . . 719

IA responsable **720**

Fundamental Limits of Membership Inference Attacks on Machine Learning Models, Aubinais Eric . . . 720

IA, durabilité et développement régional : Une approche par clusters des transitions numérique et écologique, Cafieri Simona . . . 727

Fairness is in the details : Face Dataset Auditing, Lafargue Valentin *et al.* . . . 731

Anonymization of georeferenced health microdata - a glimpse of work in progress, Lenz Rainer . . . 741

Inférence causale 1 **748**

AdaptiveCausalWVAE: Certifiably Robust and Stable Causal Disentanglement, Benhamza Manal . . . 748

Inférence de graphe de différence par découverte causale non-paramétrique, Devijver Emilie . . . 755

Ensembles de variables causales pour la généralisation de distribution, Le Goff Théotime . . . 761

The Case for Time in Causal DAGs, Reisach Alexander . . . 767

Inférence causale 2 **774**

A Generalized Propensity Score Estimation Methodology for Discrete and Continuous Treatments, Angelim Vieira Felipe Lourenço *et al.* . . . 774

Sharp Bounds for Continuous-Valued Treatment Effects with Unobserved Confounders, Baitairian Jean-Baptiste . . . 782

Policy learning under constraint: seeking primary outcome maximization under adverse outcome control, Fuentes Vicente Laura . . . 789

Measuring Variable Importance in Heterogeneous Treatment Effects with Confidence, Paillard Joseph . . . 795

Choix de variable en survie causale, Voinot Charlotte . . . 802

Informed learning **808**

Conditionnement d'un processus gaussien avec des contraintes en inégalité par projection sur une famille de B-splines, Casteras Paul . . . 808

Block-Additive Gaussian Processes under Monotonicity Constraints, Deronzier Mathis . . . 815
Résolution d'un problème inverse sur des expériences laser avec modèle génératif, Haberstich Cécile . . . 821
Problèmes inverses et apprentissage informé par la physique, Jouvin Nicolas . . . 828
Prédiction de champs physiques sous contraintes linéaires, Nassouradine Mahamat Hamdan . . . 834

Modèles de dynamiques de population pour l'écologie et l'épidémiologie (Groupe Environnement) (Session du groupe) 841
Modèle mécanistico-statistique pour l'expansion de la maladie du flétrissement du frêne, Fritsch Coralie *et al.* . . . 841
Modélisation spatio-temporelle de la distribution d'un insecte vecteur, Grosdidier Marie *et al.* . . . 848
Arbocarto, outil de modelisation de la dynamique des populations des moustiques Aedes : un vecteur de dialogue entre recherche, développement, et lutte antivectorielle, Marti Renaud . . . 854

Modèles de mélange 861
Semi-parametric estimation of location-scale mixture densities, El Hasnaoui Mohamed . . . 861
A Novel Approach to PolSAR Image Classification Using the Complex Riesz Model, Kammoun Rayhan . . . 868
Classification des données d'observance d'un traitement de l'apnée du sommeil à l'aide d'un modèle de mélange de régression basé sur une chaîne de Markov, Vernet Elodie *et al.* . . . 875

Prix actuariat + Session Groupe Risques AEF 881
Modélisation et prédiction du Best Estimate par réseau de neurones génératifs, Ebrotie Kouassi Alphonse Nathan . . . 881
Efficient estimation for EV regression models of tail risks, Kratz Marie . . . 883
Systemic risk conditional on comovement, Peng Liang . . . 886
Etude asymptotique des statistiques de cluster, Wintenberger Olivier . . . 888

Prix ENSAI-SFdS et Etudes de cas 892
Liens causaux intervenant dans le processus d'adoption d'un exosquelette dans le contexte industriel, Clerc-Urmès Isabelle . . . 892
Identification du réseau de corrélations des variables potentiellement influentes pour les biosolutions, premières modélisations, Dusfour-Castan Pauline . . . 903
Bien-être des porcs en élevages alternatifs en France, Toti Vito *et al.* . . . 914
Évaluation de modèles d'apprentissage automatique pour la prévision de la survenue d'accidents du travail à partir de séries temporelles binaires, Yapi Aho . . . 915

Prix SD-SFdS et Session Enseignement 922
SARP.moodle : créer des questions en grand nombre pour Moodle avec R. Exemple d'application à l'apprentissage des statistiques, Curis Emmanuel *et al.* . . . 922
Prix SD/SFdS : Mise en place d'une base de données pour la gestion des approvisionnements en matières premières de LOEUF, Durand Anaëlle . . . 929
Le tutorat par les pairs dans l'enseignement supérieur à accès libre : preuves de son efficacité dans les universités belges francophones, Platteau Antoine . . . 930

Processus empiriques –Empirical processes, Statistique des processus **938**
Inference of Hawkes Processes with Variable Length Memory, Quayle Sacha . . . 938

Processus Gaussiens **945**
Audio Signal Denoising with Reduced Gaussian Distributions, Fradi Anis 945
Wavelet transform for large dimension Gaussian Process regression, Kerleguer Baptiste 952
Gaussian process regression and active subspace recipe for parsimonious meta-modeling, Marrel Amandine 958

Régression avancée **964**
Régularisation Ridge de la régression PLS, Castelli Luca 964
Variational autoencoder approach for inference of nonlinear mixed effect models based on ordinary differential equations, Li Zhe 971
Régression géodésique intrinsèque dans des groupes de Lie : modèle et propriétés statistiques, Nicol Florence *et al.* 977
Approximate full conformal prediction in RKHS, Razafindrakoto Davidson Lova . 983

Sélection de variables **989**
Éstimation et sélection de variables dans les modèles non linéaires à effets mixtes, Caillebotte Antoine 989
Sélection de variables par ACP supervisée : réduction des faux positifs en grande dimension, Liquet Benoit 996
Sobol-CPI : une statistique d'importance par permutation conditionnelle doublement robuste, Reyero Lobo Angel 1002

Séries temporelles **1008**
Graph-Based Weight Cascading Methods for Multisite Time Series Forecasting, Campagne Eloi 1008
Change Point Detection in Hadamard Spaces by Alternating Minimization, Kostic Anica 1014
Estimation robuste en ligne dans le cas du modèle espace-état linéaire avec covariables stochastiques, Mahoromeza Jean-Luc 1019
Statistical Analysis of Economic Factors Influencing Consumer Choice by Vehicle Engine Type in the French Automotive Market, Verdier Claire 1027

Séries temporelles – Théorie **1031**
Sur les bruits blancs faibles modulés par une chaîne de Markov cachée, Bra Jean Armel 1031
splineOP: quadratic spline compression for univariate time series, Cecchi Nicolas 1037
Echangeabilité et Facteurs Faibles, Hallin Marc 1040
On self-excited threshold bivariate Poisson integer-valued GARCH models, Sadoun Mohamed Djemaa 1043

Session organisée par la Société Francophone de Classification **1048**
Une Approche Topologique de l'Analyse discriminante, Abdesselam Rafik 1048
Sur les structures de clustering multi-niveau définies par des convexités d'intervalle. Applications à la hiérarchie du lien simple, Bertrand Patrice 1058
Une approche divisive et interprétable de réduction de dimension pour la prédiction de la RUL, Ndao Mouhamadou Lamine 1064

Systèmes de Classes & Dissimilarités, Préa Pascal 1070

Stat & Sport (session du groupe) 1076

Clustering of soccer possessions using mixture of Hawkes processes with Markov marks, Amezouwui Koffi 1076

Application des Modèles de Markov dérivants à la dynamique d'acquisition des compétences, Vergne Nicolas 1083

PSOLxG : Une Nouvelle Métrique pour Décoder la Performance des Gardiens et Attaquants, Verronneau Rémi 1092

Statistique appliquée 1098

Vers l'automatisation de la synthèse de catalyseurs grâce à la spectroscopie Raman in situ en temps réel couplée à l'intelligence artificielle, Dubois Clara 1098

Des modes d'évolution culturelle contrastés : Langues kra-dai et technologies de tissage, Kopp Emma 1107

Contribution à l'automatisation de l'analyse de mise en page des manuscrits médiévaux : mise en place d'un test statistique pour décider si une page est vierge, Lebec Pierre 1114

Statistique appliquée à la biologie-santé 1121

Apprentissage de l'espace acoustique des pleurs de bébé, Bonafos Guillem 1121

Méthode ADD MALTS pour l'inférence causale avec outcome distributionnel : évaluation d'un programme d'hypertension sur tensiomètres connectés Withings, Chabassier Agathe 1128

Génération hiérarchique de données de comptage avec PLN-Tree: augmentation de données en microbiologie, Chaussard Alexandre *et al.* 1138

A Model-Based Clustering approach for toxicity assessment using Cell Painting data, Grigoryan Mariam 1146

ToMATo clustering algorithm for neuronal spike sorting, Martineau Louise 1152

Statistique appliquée à la consommation énergétique 1158

Explainability of electricity demand forecasting models: a Shapley value approach for positive component decomposition, Brégère Margaux *et al.* 1158

Prédiction de la consommation énergétique des trains en stationnement : Une approche économétrique basée sur le modèle PVAR, Kala Nathanaël 1165

Modélisation et simulation des défaillances dans le réseau de distribution électrique de la cie : cas du poste de la riviera, Otogo Ndoutoume Gaetan 1171

Statistique appliquée, Statistique des processus 1180

Processus de Hawkes pour évaluer le risque de tavelure du pommier, Adamczyk-Chauvat Katarzyna 1180

Statistique Bayésienne (Session du groupe) 1187

Learnings from a large simulation study to optimize Bayesian trials in small populations, Amzal Billy 1187

Estimating ratios of normalizing constants using stochastic approximation: the SARIS algorithm, Guédon Tom 1189

Calage bayésien avec entrées incertaines, application à un modèle de sas de dépressurisation, Keller Merlin 1197

Statistique bayésienne appliquée à la biologie **1207**
Inférence bayésienne de processus stochastiques intégrés et application à la phylogéographie virale, Bastide Paul . . . 1207
Modélisation de l'évolution de la présence d'une infection par un modèle bayésien de Markov caché. Application à la surveillance de la BVD en France, Galharret Jean-Michel . . . 1214
Bayesian estimation of brain functional connectivity networks, Jiang Yiye . . . 1220
Inférence de réseaux phylogénétiques par ABC-Forêts Aléatoires (ABC-RF) : application au processus de domestication du riz, Rabier Charles-Elie . . . 1226

Statistique bayésienne computationnelle **1232**
Inférence variationnelle pour l'approximation de priors de référence à l'aide de réseaux de neurones, Baillie Nils . . . 1232
Approximate Bayesian Computation with Deep Learning and Conformal prediction, Baragatti Meili . . . 1239
Améliorer la Simulation Based Inférence grâce à l'Approximate Bayesian Computation, Luciano Antoine . . . 1245
Iterated forward scheme to construct proposals for sequential Monte Carlo algorithms, Procope-Mamert Sylvain . . . 1250

Statistique biomédicale **1256**
Potential Responses and the Order of Patient Inclusion in Early-Phase Sequential Trials., Akouba Meliha . . . 1256
Un cadre pour générer des données tabulaires synthétiques réalistes dans le contexte d'un essai contrôlé randomisé, Petrakos Niki Z. . . . 1263
Score tests for non-proportional hazards in single-arm clinical trial with time-to-event endpoint, Szurewsky Chloé . . . 1272

Statistique computationnelle **1280**
Classifieur Bayésien Discret Robuste en cas de Distributions Bruitées des Covariables, Gilet Cyprien . . . 1280
Un Nouvel Algorithme de Type Full AdaGrad, Lu Wei . . . 1287
Détection de Ruptures Multivarié en Ligne : Liens avec la Géométrie Computationnelle, Rigaill Guillem . . . 1294
DUST: Duality-Based Pruning Methods For Exact Multiple Change Point Detection, Runge Vincent . . . 1299

Statistique des processus **1304**
Estimation adaptative du coefficient de diffusion dans les EDS, Huang Yichuan . 1304
Algorithme de Classification ERM-Lasso pour les Processus de Hawkes Multivariés, Lacoste Romain Edmond . . . 1311
A subsampling scheme for the estimation of Hawkes processes through spectral theory, Martinez Herrera Miguel . . . 1317
Testing procedures based on maximum likelihood estimation for Marked Hawkes processes, Sadeler Perrin Maya . . . 1323

Statistique et sport 1 **1329**
Modèle prédictif pour les performances de course à pied avec données fonctionnelles scalaires et qualitatives., Bassac Francois . . . 1329

Caractérisation et étude des profils psychologiques existants dans un échantillon de traileurs, Chauveau Alice . . . 1336
Etude de modèles de diffusion pour la modification d'images de danse sportive, Claeys Emmanuelle . . . 1341
Rugby Union as electric consumption : Short-term performance prediction in team sports using tracking data, Odet Arnaud . . . 1350

Statistique et Sport 2 **1356**
Appréhender certains facteurs qui déterminent la valeur marchande des joueurs de football, Amara-Ouali Yvenn . . . 1356
Proposition pour détecter des momentums dans un match, Derquenne Christian . 1361
Sur la détection statistique du momentum en sport, Gerville-Réache Léo . . . 1368

Statistique mathématique **1374**
Pour détecter une rupture faible dans la moyenne conditionnelle d'un modèle hétéroscédastique non-linéaire, Aouissaoui Fatma *et al.* . . . 1374
Apprentissage par transfert dans la famille exponentielle, Bourel Mathias . . . 1381
Étude de la médiane empirique pour estimer la moyenne commune de variables aléatoires hétéroscédastiques, Louati Sirine . . . 1388
Post hoc bounds for heterogeneous tests, Perier Romain . . . 1392

Statistique mathématique pour données complexes **1398**
Processus empiriques avec données hétérogènes, Boyer Jérémy . . . 1398
Tests d'uniformité sur des espaces polonais homogènes compacts, Brécheteau Claire 1406
Estimation non-paramétrique pour des données circulaires censurées, Conanec Nicolas . . . 1412
Test d'adéquation par moments trigonométriques pour K échantillons circulaires, Maxime Boucher . . . 1418

Statistique mathématique pour données en grande dimension **1426**
Dependent default times with mixed hitting-time models, Helali Salima . . . 1426
Calibration data-dépendante pour un test à noyau non-asymptotique à deux échantillons, Lacroix Perrine . . . 1433
Asymptotic subsampling for supervised learning with reproducing kernel Hilbert spaces, Vayness Eyal . . . 1439

Statistique multivariée **1444**
Propriétés et robustesse de la profondeur scatter, Louvet Gaëtan . . . 1444
Prévision Conforme pour des données hiérarchiques, Principato Guillaume . . . 1449

Statistique non paramétrique **1455**
Estimateur récursif de la densité par noyau asymétrique sous dépendance η-faible, Adjabi Smail . . . 1455
Heavy-tailed and Horseshoe priors for regression and sparse Besov rates, Egels Paul 1462
MCD: Marginal Contrastive Discrimination for conditional density estimation, Ndiaye Aminata . . . 1470
Nonparametric estimation for an integral transform of the transition density, Schürzinger Felix . . . 1476

Statistique pour l'industrie **1482**

Algorithme génétique pour la corrélation de transfert de chaleur en ébullition par film, Beaucourt Julien . . . 1482
Modèle d'apprentissage profond explicable pour l'amélioration de la surveillance de la radioactivité alpha aéroportée., Roblin Arthur . . . 1489
Clustering methods to segregate avionic failures: A case study from the aerospace industry, Sanchez San Benito Alvaro . . . 1495
Identification of Graph Features to Improve GNN Performance, Zehavi Itai . . . 1501

Statistique pour la société **1504**
Vers une carte des élections basées sur des évaluations, Aubin Jean-Baptiste . . . 1504
Unpacking the Map of Opportunity: a multidimensional approach to the sub-national variations of social mobility in France, Gomtsyan Marina . . . 1511
Money laundering detection: a contrastive learning approach on financial time series with a transformer, Guéneau Harold . . . 1518
Comparing and Proposing Methods for Estimating Bilateral Migration Flows, Laurent Thibault . . . 1524

Statistique spatiale **1531**
Lead Prediction Error Analysis in 2D-RCAR Random Fields, Arezki Ouerdia . . . 1531
Statistique des points critiques d'un champ gaussien, Chevallier Julien . . . 1540
Estimation de la régularité d'un champ gaussien, Makhlouf Alexandre . . . 1545
Spatio-temporal random fields on meshed surfaces defined from advection-diffusion SPDEs, Pereira Mike . . . 1550
Deep kernel learning for geostatistics, Romary Thomas . . . 1556

Traitement automatique du langage naturel (Groupe MALIA) **1561**
SigBERT: Combiner des compte-rendus médicaux de patients et la théorie des Signatures de chemins pour la Prédiction de Survie, Minchella Paul . . . 1561
Large Language Models as Markov Chains, Odonnat Ambroise . . . 1568
Data Complexity is not Data uncertainty, Sinha Aman . . . 1576

Transport optimal **1577**
Procrustes-Wasserstein for isometric invariant interpolations, Adamo Davide . . . 1577
Optimal Transport-Based Domain Adaptation for Rotated Linear Regression, Brian Britos . . . 1583
Aligning Embeddings and Geometric Random Graphs: Informational Results and Computational Approaches for the Procrustes-Wasserstein Problem, Ganassali Luca 1589
Prédiction par krigeage de mesures de probabilité. Application en simulation numérique pour les études de sûreté nucléaire, Gossard Florian . . . 1596
Barycentres de Mesures pour Coûts de Transport Génériques, Tanguy Eloi . . . 1603

Index des auteurs **1609**

Conférence Lucien Le Cam

Bayesian semi-parametric : a curse or a blessing

Judith Rousseau*[1]

[1]CEntre de REcherches en MAthématiques de la DEcision – Université Paris Dauphine-PSL, Centre National de la Recherche Scientifique, Centre National de la Recherche Scientifique : UMR7534 / URA749 – France

Résumé

In this talk I will first review the advances on semiparametric Bayesian inference, focusing on the Bernstein von Mises property (BvM) which allows to construct credible regions that are confidence regions. While Bayesian semi parametric methods are natural methods in that they handle the nuisance parameters by integrating them out; proving the BvM property is a rather difficult task which remains an open problem in many (even simple) cases. Recently two different tricks have been introduced to obtain reliable Bayesian inference in models where the data are iid and the efficient influence function has an explicit expression. The first is a *contamination* of the prior and the second is a perturbation (or correction) of the posterior. In this talk I will discuss two cases dealing with the more complicated cases where the efficient influence function is not explicit, naming the case of regression models with missing covariates and the Hawkes model.

These results are joint works with Stefan Franssen and Mael Duverger

Mots-Clés: Bayesian statistics, asymptotic statistics, semi, parametric inference, Bernstein von Mises property, missing data

*Intervenant

Prix Laplace

Small noise statistics

Sara Van De Geer*[1]

[1]ETH Zürich – Suède

Résumé

We consider regression and classification with small noise. When there is little noise the statistical problem is easier. But small noise also poses theoretical challenges. This is exemplified in some
nonparametric regression problems with random design, and in (high-dimensional) logistic regression.

Mots-Clés: nonparametric regression, logistic regression, signal, to, noise ratio

*Intervenant

Prix du Docteur Norbert MARX

Nonlinear multilevel joint model for individual lesion kinetics and survival to characterise intra-individual heterogeneity in patients with advanced cancer

Marion Kerioui*[1,2]

[1]MRC Biostatistics Unit, University of Cambridge – Royaume-Uni
[2]Université Paris Cité, INSERM, IAMEF-75018, Paris, France – Centre de Recherche Inserm – France

Résumé

In oncology, follow-up of patients with advanced disease commonly relies on tumour burden, calculated using the sum of the longest diameters (SLD) of the target lesions, a measure that lumps all lesions together and ignores intra-patient heterogeneity. Due to their complex mechanism of action, immunotherapies are associated with increased variability in response to treatment, and higher risk of observing divergent kinetic profiles within the same patient. In this project, we aimed to assess the benefit of modelling target lesions' kinetics to better understand the variability in response to immunotherapies and its implications for patients' outcome. We developed a Bayesian hierarchical joint model that disentangle different sources of variability in lesions' kinetics, and characterise their association with the risk of death. The model parameters were estimated on 900 patients from a phase III randomized trial evaluating programmed death-ligand 1 checkpoint inhibitor atezolizumab versus chemotherapy in patients with second-line metastatic urothelial carcinoma (IMvigor211). Relying on a tumour growth inhibition model, we estimated that between-lesions variability account for between 12% and 78% of the total variability during chemotherapy. We found a similar contribution of the lesion level to the total variability in patients treated with atezolizumab, except for the durability of treatment-induced tumour shrinkage, for which between-lesions variability was markedly larger than during chemotherapy (40% vs 12%, respectively). We showed that kinetics of lesions located in the liver and in the bladder had twice as much impact on the instantaneous risk of death compared to those located in the lung or the lymph nodes. We then aimed to assess the ability of lesions' kinetics to predict the individual risk of death, in comparison to the usual marker, the SLD. Using an individual dynamic prediction approach, we showed that describing individual trajectories of all target lesions provided a better prediction of patients most at-risk of death than a model relying solely on the SLD. With our proposed model, we pave the way for a better understanding of mechanisms of response associated to immunotherapies and better monitoring of cancer patients, to quickly adapt therapeutic strategies in case of disease progression.

Mots-Clés: Advanced cancer, Bayesian inference, Joint model, Multilevel modeling, nonlinear mixed, effects model

*Intervenant

Sessions prix étudiants

Modélisation et prédiction du Best Estimate par réseau de neurones génératifs

Kouassi Alphonse Nathan Ebrotie*[1]

[1]EURIA – Prix de Statistiques pour l'Actuariat 2025 – France

Résumé

La réglementation Solvabilité ii (Sii) impose aux assureurs un certain nombre de contraintes dont l'évaluation de leurs provisions techniques de la façon la plus juste possible, ces provisions ayant une composante dite Best Estimate (BE). Un autre enjeu de Sii, concerne l'estimation prospective de certaines métriques comme le ratio de Solvabilité dans le cadre de l'ORSA ou pour des allocations stratégiques d'actifs. Afin de respecter ces exigences du régulateur, les assureurs utilisent généralement des modèles d'Asset and Liability Management (ALM) afin d'effectuer des projections des différents flux financiers en fonction de scénarios économiques. À partir de ces projections, les différents éléments d'intérêts (BE, ratio, etc.) peuvent être obtenus en utilisant des méthodes basées sur des simulations comme l'approche classique de Simulations dans les Simulations (SdS). Toutefois, en pratique, cette méthode fondée sur le Monte-Carlo est très couteuse en temps de calcul. C'est pour cette raison que des techniques de réduction de temps de calcul ont été développées. Dans la littérature actuarielle, plusieurs mémoires se sont attardés sur la mise en place de ces métamodèles pour des fins identiques à ceux évoqués plus haut, mais adoptent des approches alternatives qui ne se généralisent pas souvent et ne permettent pas parfois de capturer efficacement les distributions temporelles des éléments d'intérêts, sachant que cette distribution peut être utile pour des problématiques d'estimations prospectives. Avec l'avènement d'outils comme ChatGPT, DALL-E, Midjourney, les modèles génératifs font parler d'eux de façon spectaculaire, et laissent entrevoir de nouvelles opportunités en termes de modélisation au vu des capacités des modèles sous-jacents. C'est dans la perspective de tirer profit des possibilités qu'offrent ces modèles génératifs dans des tâches de modélisation que ce mémoire s'inscrit. Notre étude se focalise sur l'utilisation des modèles génératifs afin d'approximer le Best Estimate. Cette estimation s'intègre également dans un cadre prospectif et peut permettre de réaliser des anticipations aussi bien sur le BE que sur l'évolution de métriques comme le SCR et le ratio de Solvabilité.

Mots-Clés: Solvabilité ii, Best Estimate, proxy, réseaux de neurones génératifs.

*Intervenant

Bien-être des porcs en élevages alternatifs en France

Vito Toti*[1], Maxime Tio*[1], and Nathan Saraiva[1]

[1]Ecole Nationale de la Statistique et de l'Analyse de l'Information [Bruz] – Ensai, Ecole Nationale de la Statistique et de l'Analyse de l'Information – France

Résumé

Les élevages alternatifs de porcs connaissent un essor significatif en France et répondent à la demande croissante des ménages soucieux de la qualité et de l'origine des produits qu'ils consomment. Toutefois, ces systèmes ont des contraintes différentes par rapport aux élevages intensifs, et comportent des risques pour la santé animale en raison d'une exposition accrue aux maladies et aux agents pathogènes. Ces conditions mettent en péril le bien-être des porcs et peuvent causer des ravages considérables chez les animaux. C'est dans ce contexte que le programme PIGAL, au sein duquel nous nous inscrivons, a pour objectif d'étudier les facteurs qui contribuent le plus au bien-être dans ces environnements. Nous avons procédé au traitement et à l'analyse d'échantillons recoupant différents types d'élevages, à savoir les élevages de truies reproductrices (Naisseurs), les élevages de porcs en croissance (Engraisseurs) et les élevages combinant les deux activités (Naisseurs-Engraisseurs). Compte tenu du grand nombre de variables explicatives et du nombre limité d'élevages étudiés, nous avons utilisé des méthodes adaptées à ce type de données, telles que la régression PLS multiblocs et la régression Elastic Net. Nos travaux ont mis en évidence une variété kaléidoscopique de facteurs de bien-être animal, regroupés au sein de thématiques comprenant l'alimentation, l'hygiène, le logement ou encore la santé. Ces résultats serviront à élaborer des recommandations destinées aux éleveurs et aux pouvoirs publics, afin d'éclairer leurs décisions et améliorer ainsi le bien-être des porcs dans les élevages alternatifs.

Mots-Clés: Projet, Statistique, Porcs, Elevages, ENSAI, Bien, être

*Intervenant

Prix SD/SFdS : Mise en place d'une base de données pour la gestion des approvisionnements en matières premières de LOEUF

Anaëlle Durand*[1]

[1]Université Bretagne Sud – Master Science des données – France

Résumé

LOEUF est une entreprise grandissante qui a un besoin de fiabiliser ses données et ses prévisions
en matière d'approvisionnement d'œufs.

L'objectif de l'alternance est de répondre à ce besoin en modélisant la ponte avec un modèle additif

généralisé afin d'avoir des estimations précises. Ce modèle donnera lieu à une table dans une base de

données qui permettra, à partir du parc actuel de LOEUF, de prédire le nombre d'œufs entrant dans l'usine

pour chaque semaine.

Ce projet nécessite plusieurs acteurs qui interviennent à différents moments du projet. Nous avons

d'abord la partie collecte des données dans Arcadia géré par le pôle Amont. La partie " flux entre Cognos

et Arcadia " géré par LOEUF et le siège LDC et la dernière partie " construction de tableaux de bords géré

par LOEUF.

Les prévisions actuelles sont suffisantes pour le moment, mais elles prennent du temps à être réalisé

et sont de moins en moins fiable avec l'arrivée de nouveaux élevages. La création de l'outil devient
indispensable avec la croissance de l'entreprise.

*Intervenant

Sessions invitées semi-plénières

Bayesian deep learning, overview and challenges

Julyan Arbel*[1]

[1]MISTIS (INRIA Grenoble Rhône-Alpes / LJK Laboratoire Jean Kuntzmann) – Laboratoire Jean Kuntzmann, INRIA – France

Résumé

Bayesian deep learning is appealing as it combines the coherence and natural uncertainty quantification of the Bayesian paradigm together with the expressivity and compositional flexibility of deep neural networks. It has its roots in pioneering work by Radford Neal and David Mackay in the 1990s on Bayesian neural networks. Its strengths lie in principled uncertainty quantification, improved data efficiency, and adaptability, making it impactful in safety-critical fields like healthcare and autonomous systems. In this talk I will provide an overview of Bayesian deep learning and discuss some of the key challenges the field faces in addressing modern machine learning problems.

Mots-Clés: Bayesian statistics, neural networks

*Intervenant

Processus de Markov déterministes par morceaux et Processus Décisionnels Markoviens, comment les combiner pour mieux soigner.

Alice Cleynen*[1]

[1]CNRS – CNRS, CNRS : UMR8568, CNRS, CNRS : UMR6074, CNRS, CNRS : UMR5593, CNRS : ERL3189, CNRS : UMR7104, CNRS : UMR5244, CNRS : UMR2205, CNRS : UPR8241, CNRS, CNRS : UMR5554, CNRS : UMR5274, CNRS : UMR5493, CNRS : UMR7199, CNRS : UMR8184, CNRS : UMR7141, CNRS : UMR5251, CNRS : UMRTemps8066, CNRS : FR550, CNRS : UMR5127, CNRS : UMRSETEMoulis, CNRS : UMR9189, CNRS : UMR7190, CNRS : UMR00, CNRS : UMR8079, CNRS : UMR5138, CNRS : UMR6004, CNRS : UMR8234, CNRS : UMR5249, CNRS : UMR5109, CNRS : UMR8503 – France

Résumé

Dans les maladies de longue durée comme le cancer, les patients alternent entre des phases de rémission et de rechute, et sont suivis dans le temps au moyen de contrôles non invasifs tels que des prises de sang. À partir de ces mesures indirectes et bruitées de certains marqueurs de la maladie, les praticiens doivent décider de l'allocation des traitements, parfois avec peu de connaissances sur la dynamique du processus (par exemple, l'agressivité de la rechute), qui peut varier d'un patient à l'autre. Ce problème peut être modélisé dans le cadre du contrôle impulsionnel de processus de Markov déterministes par morceaux (PDMP) à paramètres mal connus et partiellement observés, dont la résolution est encore une question ouverte.

Dans cet exposé, je retracerai l'histoire de la résolution de tels problèmes depuis la brique la plus simple, le processus décisionnel de Markov (MDP) en temps discret, parfaitement observé et parfaitement connu, jusqu'à des contributions récentes sur le contrôle impulsionnel des PDMP, toujours illustrée par notre problème médical. Je tisserai des liens entre MDP et PDMP qui ouvrent des pistes prometteuses pour la résolution des problèmes les plus complexes.

Cette présentation est le fruit d'un travail en collaboration avec Benoîte de Saporta, Orlane Rossini, Régis Sabbadin et Amélie Vernay.

Mots-Clés: PDMP, MDP, Contrôle, Programmation dynamique, Allocation de traitement

*Intervenant

Spatio-temporal models that combines different type of extremal dependencies

Carlo Gaetan*[1]

[1]Dipartimento di Scienze Ambientali, Informatica e Statistica [Venezia] (DAIS) – San Giobbe, Cannaregio 873 I-30121 Venezia, Italie

Résumé

The analysis of spatial extremes has received a significant boost because many natural extreme hazards, such as heat waves, heavy rain and snowfall, storm surges, and windstorms, have a spatial extent.
Much effort has been devoted to proposing spatial models based directly on extreme value theory, such as max-stable processes and Pareto processes. However, by construction, two salient properties of these classes of asymptotic models (maximum and threshold stability) are rarely satisfied by the environmental or climatological data available at the observed extremal levels. For this reason, there have been several recent attempts in the literature to construct sub-asymptotic models for spatial extremes based on the exceedances of high thresholds, combining tail flexibility with computational tractability.

When studying the extremal dependence of threshold exceedances, a first distinction is made between asymptotic dependence and asymptotic independence. These two limiting forms are characterized, respectively, by persistence of dependence
or convergence to independence among observations exceeding asymptotically increasing thresholds.

Modeling the spatio-temporal extremes adds an additional layer of complexity. Indeed, different forms of asymptotic dependence may arise in the spatial and temporal domains.

In this presentation, we first review some of these issues and then present a class of spatio-temporal models
for threshold exceedances that embeds different forms of limiting dependence according
to the different dimensions considered, spatial or temporal. We provide a parametrization that is able to distinguish between asymptotic dependence and independence, both in space and in time, and this fact can be exploited for parametric inference.

Since likelihood-based inference from observed exceedances is not feasible, we propose to estimate the parameters of the models by combining simulations and neural network learning.
We exemplify the method using transformed Gaussian random processes, since this class of models provides a convenient way to specify spatio-temporal dependence and simulate spatio-temporal observations.
We conclude the talk with a real data example.

*Intervenant

This presentation is the result of collaboration with Paola Bortot (University of Bologna, Italy) and Lorenzo Dell'Oro (University of Padova, Italy).

Mots-Clés: Asymptotic dependence, asymptotic independence, Gaussian random processes, neural networks

Algorithmes efficaces pour la détection optimale des ruptures dans les données séquentielles

Toby Hocking*[1]

[1]Université de Sherbrooke – Canada

Résumé

Time and memory efficiency are important features of statistical software. When several different implementations of a given algorithm are available, benchmarking can be used to determine which is most efficient for a given task. Typical benchmarking tools provide support for comparing time/memory usage of different computations for a data set of a given size N. We present the atime package, which provides asymptotic benchmarking, meaning that time and memory (as well as other quantities) can be easily measured for a sequence of increasing data sizes. Asymptotic measurement allows estimation of complexity classes (big-O notation), and provides a robust new method for testing the performance of R packages. We discuss how atime has been used to improve efficiency in base R and data.table.

Mots-Clés: données séquentielles, optimisation, détection de ruptures, programmation dynamique, algorithme efficace, statistique computationnelle

*Intervenant

Identification et caractérisation de la variabilité individuelle au sein d'une population à partir d'un modèle mécaniste et d'effets mixtes. Applications en amélioration variétale

Estelle Kuhn*[1]

[1]Université Paris Saclay – INRAE – France

Résumé

Dans le contexte de changement climatique, les évolutions des besoins en alimentation et de la pression sur les ressources imposent à l'agriculture de revoir ses méthodes de production. Il devient essentiel de développer des systèmes de culture et d'élevage innovants, durables et résilients. La diversité génétique des plantes et des animaux est un atout majeur pour créer ces nouveaux systèmes. Pour capitaliser efficacement sur cette diversité, il est nécessaire de mieux comprendre l'interaction entre une plante ou un animal et son environnement. Quelles variétés cultivées ou races élevées actuellement pourront s'adapter au changement climatique ? Peut-on identifier des leviers d'action biologiques afin de sélectionner des variétés mieux adaptées aux nouveaux objectifs ?

Ces dernières années, des modèles mathématiques décrivant les processus biologiques de développement et de production des plantes ou des animaux ont été développés à partir des connaissances d'experts. Ils intègrent les conditions environnementales et dépendent de paramètres qui sont souvent mécanistes, représentant une quantité biologique comme un taux d'absorption en azote par exemple. Pour une même espèce, la valeur de ces paramètres varie possiblement pour différentes variétés ou génotypes. Cette variabilité représente en partie celle des mécanismes biologiques sous-jacents. Il est ainsi primordial de bien caractériser la variabilité de ces paramètres au niveau de la population de variétés pour identifier des leviers pertinents pour la sélection. En particulier il est important d'identifier les paramètres portant de la variabilité et de la caractériser à l'aide de variables descriptives des génotypes, par exemple des marqueurs génétiques de grande dimension.

Dans cet exposé, je présenterai différentes méthodes hybrides combinant modélisation mécaniste et approche populationnelle et discuterai leurs avantages et limites. Je détaillerai en particulier des résultats statistiques récents d'estimation, de test des composantes de la variance et de sélection de variables en grande dimension dans des modèles à effets mixtes non linéaires, permettant d'identifier et de caractériser la variabilité individuelle des paramètres mécanistes.

Cet exposé inclut des résultats de travaux réalisés en collaboration avec Charlotte Baey, Antoine Caillebotte, Maud Delattre, Tom Guédon, Jean-Benoist Leger, Sarah Lemler.

*Intervenant

Mots-Clés: Modèles à effets mixtes, modélisation mécaniste, test des composantes de la variance, sélection de variables, grande dimension

Modèles de mélange nonparamétriques à variables latentes dépendantes

Claire Lacour*[1]

[1]Laboratoire Analyse et de Mathématiques Appliquées – Université Gustave Eiffel – France

Résumé

The talk will be in French with slides in English.
Dans cet exposé on s'intéressera aux modèles de mélange finis nonparamétriques, en particulier quand les variables latentes étiquetant la population ne sont pas indépendantes mais markoviennes. On introduira une extension des modèles de Markov cachés, dans laquelle les variables observées ne sont plus indépendantes, même conditionnellement aux variables latentes. On cherchera des estimateurs dans le cas particulier d'une population à dynamique markovienne connue contaminée aléatoirement par une population à dynamique indépendante inconnue.

Mots-Clés: Modèles de mélange, Estimation nonparamétrique, Chaînes de Markov

*Intervenant

Méthodes d'ensembles en apprentissage

Pierre-Alexandre Mattei*[1,2]

[1]Equipe Maasai, Inria Sophia Antipolis – Institut National de Recherche en Informatique et en Automatique – France

[2]Laboratoire Jean Alexandre Dieudonné – Université Nice Sophia Antipolis (1965 - 2019), Centre National de la Recherche Scientifique, Université Côte d'Azur – France

Résumé

Les méthodes d'ensembles combinent les prédictions de différents modèles d'apprentissage statistique. Leurs représentants les plus fameux sont les forêts aléatoires ou les ensembles de réseaux de neurones.
Cet exposé tournera autour de la question "combien de modèles dois-je agréger ?". On verra que la réponse dépend de la mesure de performance choisie. En particulier, dans le cas de pertes convexes (comme l'entropie croisée en classification ou l'erreur quadratique en régression), l'erreur est fonction décroissante du nombre de modèles. Dans le cas de pertes non-convexes (comme l'erreur de classification), les choses sont plus nuancées, et l'erreur peut parfois être non-monotone, même asymptotiquement.
Ces résultats seront illustrés par des exemples d'ensembles de réseaux de neurones. Ce travail sera notamment basé sur la prépublication https://arxiv.org/abs/2311.17885, réalisée avec Damien Garreau (Julius-Maximilians-Universität Würzburg).

Mots-Clés: Ensembles, machine learning, random forests

*Intervenant

Statistical learning viewpoints on extreme value analysis

Anne Sabourin*[1,2]

[1]Mathématiques Appliquées Paris 5 – Institut National des Sciences Mathématiques et de leurs Interactions - CNRS Mathématiques, Centre National de la Recherche Scientifique, Université Paris Cité – France

[2]Télécom Paris – Institut Mines-Télécom [Paris], Institut Polytechnique de Paris – France

Résumé

Recent years have seen significant theoretical and methodological developments at the intersection of Extreme Value Analysis (EVA), machine learning, and statistical learning. In this talk, I will discuss recently introduced frameworks for supervised learning with extreme covariates and for dimension reduction in extreme-value settings. I will also highlight key shared proof techniques, particularly those concerning the deviations of empirical processes indexed by low-probability classes, which may be leveraged for settings outside EVA such as imbalanced classification. The presentation will be primarily based on the following works:
- Clémençon, S., Huet, N., & Sabourin, A. (2024). Regular variation in Hilbert spaces and principal component analysis for functional extremes. Stochastic Processes and their Applications, 174, 104375.

- Aghbalou, A., Bertail, P., Portier, F., & Sabourin, A. (2024). Cross-validation on extreme regions. Extremes, 27(4), 505-555.

- Aghbalou, A., Portier, F, & Sabourin, A. (2024) Sharp error bounds for imbalanced classification: how many examples in the minority class? Proceedings of AISTATS.
- Clémençon, S., Jalalzai, H., Lhaut, S., Sabourin, A., & Segers, J. (2023). Concentration bounds for the empirical angular measure with statistical learning applications. Bernoulli, 29(4), 2797-2827.

Mots-Clés: Extreme values, Statistical Learning, High dimension

*Intervenant

Two problems in smooth regression: shape constraint and auto-correlation

Wood Simon*[1]

[1]University of Edinburgh – Royaume-Uni

Résumé

Smooth regression models (GAM, GAMM, GAMLSS etc) are widely used in statistical applications, but two practical problems are quite common. The first is that there may be short range auto-correlation present in the data that is not of direct interest, but which will degrade inference if ignored. This is particularly common with data for which space or time are important. The second is that the relationship between some predictors and the response may be subject to known shape constraints (monotonicity, convexity etc), which it would be advantageous to include in the model formulation. This talk will discuss straightforward solutions to both issues that integrate well with widely used sooth regression frameworks and software, illustrated with practical examples.

Mots-Clés: generalized additive model, neighbourhood cross validation, shape constrained smoothing

*Intervenant

Pricing Insurance Contracts and Managing Financial Balances within an Insurance Portfolio

Julien Trufin*[1]

[1]Service de Physique Théorique [Bruxelles] (SPT-ULB) – Belgique

Résumé

Machine learning techniques provide actuaries with predictors exhibiting high correlation with claim frequencies and severities. However, these predictors generally fail to achieve financial equilibrium and thus do not qualify as pure premiums. Autocalibration effectively addresses this issue since it ensures that every group of policyholders paying the same premium is on average self-financing. This talk proposes to look at recent results concerning autocalibration. In particular, we present a new characterization of autocalibration which enables to identify whether a pure premium model is autocalibrated or not, we study a method (called balance correction) for obtaining an autocalibrated pure premium from any regression model, we highlight the effect of balance correction on resulting pure premiums, and finally we go through some performances criteria that are particularly relevant for autocalibrated pure premiums.

Mots-Clés: Regression models, Balance properties, autocalibration, performance criteria

*Intervenant

Hommage à Pierre Cazes

Pierre Cazes : l'éditeur, le statisticien, l'homme

Xavier Bry*[1], Christian Derquenne*[2], and Gilbert Saporta*†[3]

[1]Institut Montpelliérain Alexander Grothendieck – Centre National de la Recherche Scientifique, Université de Montpellier – France

[2]Société française de statistique – SFDS – France

[3]Centre d'Etude et De Recherche en Informatique et Communication (CEDRIC) – Conservatoire National des Arts et Métiers (CNAM) – 292 Rue St Martin FR-75141 Paris Cedex 03, France

Résumé

Pierre Cazes, membre d'honneur de la SFdS, est décédé le 12 novembre 2024 à Paris. Il aurait eu 84 ans à quelques jours près.
Pierre était un pilier de notre communauté, de l'ASU puis de la SFdS, soucieux de promouvoir la statistique sous ses différentes formes. Il fut rédacteur en chef de la *Revue de Statistique Appliquée* de 1979 à 2006 jusqu'à sa fusion avec le *Journal de la SFdS* qui succédait au *Journal de la Société de Statistique de Paris*. En tant que lecteur, relecteur ou rapporteur, sa méticulosité était légendaire, mais elle s'exerçait toujours dans un esprit bienveillant pour faire progresser l'auteur.

Connu pour ses travaux en analyse des données autour de Jean-Paul Benzécri, membre du comité de rédaction des *Cahiers de l'Analyse des Données*, auteur de nombreuses publications en régression, en analyse factorielle, en classification, sur la décomposition en blocs de tableaux, l'analyse de la non-réponse et le codage de données, Pierre avait une culture statistique exceptionnelle s'étendant bien au-delà des thèmes précédents.

Tant à l'ISUP qu'à l'Université Paris-Dauphine où il exerça jusqu'en 2010, il forma des centaines d'étudiants et encadra de nombreuses thèses.

Pierre avait été élu membre de l'Institut International de Statistique en 1985 et il fut Président de la Société Francophone de Classification.

A travers des témoignages écrits et oraux d'anciens collègues et doctorants, cette session spéciale retracera la carrière de Pierre Cazes, ses contributions scientifiques, son rôle d'éditeur et l'homme qu'il était avec ses immenses qualités alliant modestie, simplicité et disponibilité, à une grande rigueur intellectuelle.

Mots-Clés: Cazes, Revue de statistique appliquée, analyse des données

*Intervenant

†Auteur correspondant: gilbert.saporta@cnam.fr

FENStatS

Workshop on FENStatS Accreditation of Statisticians

Magnus Pettersson*†[1]

[1]FenStatS – Suède

Résumé

Statisticians are working in important and highly competitive fields of applications - However, anyone can call themselves statisticians! To improve the quality of statistical work, ethics and quality, FENStatS has started offering accreditation of statisticians in cooperation with its national member associations (including SFdS). Meet FENStatS accreditation committee and learn more how you can be the next accredited statistician!

Mots-Clés: Accréditation, FenStatS

*Intervenant

†Auteur correspondant: magnus.pettersson@statistikkonsulterna.se

Session Groupe Jeunes

Notre rapport à la compétition dans l'enseignement supérieur et la recherche : avons-nous besoin de champions ?

Guillaume Chennetier [1] & Maya Sadeler-Perrin [2] & Eloi Tanguy [3]

[1] *CERMICS, École des Ponts, Institut Polytechnique de Paris, Marne-la-Vallée, France.*
[2] *Laboratoire Jean Kuntzmann, Université Grenoble Alpes, Saint-Martin-d'Heres, France.*
[3] *Université Paris Cité, CNRS, MAP5, F-75006 Paris, France.*

guillaume.chennetier@enpc.fr ; maya.sadeler@univ-grenoble-alpes.fr ; eloi.tanguy@u-paris.fr ;

Résumé. Le Groupe Jeunes de la SFdS s'engage depuis plusieurs années maintenant à créer un espace de discussion nécessaire pour sensibiliser la communauté académique à des sujets difficiles, comme la santé mentale des jeunes chercheur.e.s ou l'éthique dans la recherche. La session spéciale du groupe Jeunes prend place chaque année aux Journées de Statistique et a pour but de promouvoir l'échange et de proposer un cadre de réflexion sain et constructif autour d'une thématique choisie, non technique, sous la forme d'une session interactive entre les intervenants et le public. Nous proposons cette année le thème de la compétition dans l'enseignement supérieur et la recherche sous toutes ses facettes : pression à la publication, campagnes de recrutement, course aux prix, sélection des étudiants, politiques de financement, mise en concurrence des établissements, etc. Nous essaierons de décortiquer ensemble les effets de cette compétition sur l'organisation de notre travail et de questionner son bien fondé. Critique-t-on trop facilement les ressorts d'une saine émulation, seule voie vers l'excellence académique ? Est-elle un inconfort nécessaire pour faire face aux réalités économiques et à la concurrence internationale ? Quel tri opère-t-elle vraiment ? Peut-on totalement contester le besoin d'évaluation ? Quelle place pour l'ambition individuelle ou collective ? Que concède-t-on en priorisant le bien-être au travail pour toutes et tous ?

Mots-clés. Compétition, Sélection, Performance, Classements, Évaluation, Biais, Système de publications scientifiques.

Abstract. The SFdS Youth Group has been committed for several years now to creating a necessary space for discussion to raise awareness within the academic community about challenging topics, such as the mental health of young researchers or ethics in research. The youth session takes place annually at the Journées de Statistique to promote exchange and offer a framework for healthy and constructive reflection on a chosen non-technical theme, in the form of an interactive session between the speakers and the audience. This year, we propose the theme of competition in higher education and research in all its facets : pressure to publish, recruitment campaigns, the race for awards,

student selection, funding policies, competition among institutions, etc. Together, we will attempt to dissect the effects of this competition on the organization of our work and question its validity. Do we too easily criticize the mechanisms of healthy emulation, the only path to academic excellence ? Is it a necessary discomfort to face economic realities and international competition ? What kind of selection does it really enforce ? Can we completely contest the need for evaluation ? What role is there for individual or collective ambition ? What do we concede by prioritizing well-being at work for all ?

Keywords. Competition, Selection, Performance, Ranking, Assessment, Bias, Scientific publishing system.

Apprentissage & données énergétiques

Spatio-temporal clustering and reconciliation for regional electricity demand forecasting

Margaux Brégère [1] & Raffaele Mattera [2]

[1] *LPSM, Sorbonne University, Paris, France and EDF R&D, Lab Paris-Saclay, France, margaux.bregere@edf.fr*

[2] *Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Caserta, Italy*

Résumé. Une approche en trois étapes est proposée pour prévoir les séries temporelles de la demande d'électricité de plusieurs zones appartenant à une même région. Tout d'abord, les séries temporelles des différentes zones sont agrégées sur la base d'un *clustering* spatio-temporel. Ceci permet de construire une hiérarchie à trois niveaux (zones / *clusters* / région); la consommation de la région étant la somme de celles des *clusters*, elles-mêmes sommes de celles des zones membres des *clusters*. Ensuite, des prévisions sont générées pour chacunes des séries temporelles à l'aide de modèles additifs généralisés. Enfin, les prévisions sont projetées de manière optimale sur un sous-espace contraints afin de s'assurer que les prévisions finales satisfont les contraintes hiérarchiques. L'approche est testée sur les données de la consommation d'électricité de la région Alberta au Canada ; les résultats expérimentaux suggèrent que les étapes successives de *clustering* et de projection améliorent les prévisions initales à la fois au niveau agrégé (région) et au niveau désagrégé (zones).

Mots-clés. *Clustering* spatio-temporel, Réconciliation de prévisions, Prévision de consommation électrique

Abstract. A three-stage approach is proposed to forecast the electricity demand time series of several areas belonging to the same region. First, time series are aggregated on the basis of a spatio-temporal clustering approach, and a three-level hierarchy is build. Second, benchmark forecasts are generated for all series using generalized additive models. Finally, the forecasts are optimally projected onto a coherent subspace to ensure that the final forecasts satisfy the hierarchical constraints. Our approach is tested on Alberta electricity demand data; experimental results suggest that successive clustering and projection steps improve the benchmark forecasts both at the aggregate level and at the dis-aggregated level.

Keywords. Spatio-temporal clustering, Reconciliation forecasting, Electricity demand forecasting

1 Introduction

To balance the supply and demand of electricity and keep the grid operating 24 hours a day, it is crucial to predict the electricity demand as accurately as possible at various levels of aggregation. In this work, we consider local electricity demand time series (bottom-level), as well as the sum of them (top-level). We aim to forecast both. To do so, we first build a three-level hierarchy by aggregating bottom-level time series based on a spatio-temporal clusterin. Second, benchmark forecasts are independently generated for all series (bottom-level, clusters and top-level) using generalized additive models. Noticing that these time series are linked through summation constraints, the forecasts are finally optimally reconciled using the Minimum Trace reconciliation approach (MinT) introduced by Hyndman et al. (2011). Our approach is tested on Alberta electricity demand data; experimental results suggest that successive clustering and reconciliation steps improve the benchmark forecasts both at bottom and top levels.

2 Methodology

Our intention is to forecast the time series of local (bottom-level) electricity demand in m areas $\left\{(\mathrm{b}_t^i)_{t>0},\, i = 1,\ldots,m\right\}$ as well as the global demand $(\mathrm{d}_t)_{t>0}$, with $\mathrm{d}_t = \sum_{i=1}^{m} \mathrm{b}_t^i$. First, space-time series $(\mathrm{b}_t^i)_{t>0}$ are clustered into k clusters. By re-indenting the time series according the clustering such that the first n_1 belong to Cluster 1, the following n_2 to Cluster 2 and so one, we define $m_0 = 1$, and, for $\ell = 1,\ldots,k$ $m_\ell = \sum_{j=1}^{\ell} n_j$, and we create the k middle-level time series $\left\{(\mathrm{c}_t^\ell)_{t>0},\, \ell = 1,\ldots,k\right\}$ with $\mathrm{c}_t^\ell = \sum_{i=m_{\ell-1}}^{m_\ell} \mathrm{b}_t^i$, for any time step t. This induces the hierarchy represented in Figure 1. For any time step t, with $\mathbf{b}_t$ be the vector

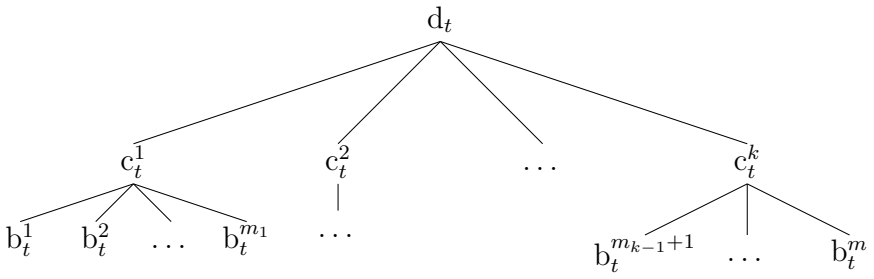


Figure 1: Representation of the three-level hierarchy induced by the clustering.

of the m bottom-level time series at time t, and $n = m + k + 1$, we define the n-dimensional vector containing all the time series $\mathbf{y}_t$: $\mathbf{y}_t = \begin{bmatrix} \mathrm{d}_t & \mathrm{c}_t^1 & \cdots & \mathrm{c}_t^k & \mathbf{b}_t^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} = \mathbf{S}\mathbf{b}_t$, where $\mathbf{S}$ is the

$n \times m$ "summing" matrix defined thanks to the clustering:

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{I}_m \end{bmatrix} \text{ with } \boldsymbol{A} = \begin{bmatrix} 1 & \dots & 1 & 1 & \dots & 1 & \dots & 1 & \dots & 1 \\ 1 & \dots & 1 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \underbrace{0 \dots 0}_{n_1} & & & \underbrace{1 \dots 1}_{n_2} & & & \dots & \underbrace{1 \dots 1}_{n_k} & & \end{bmatrix} \begin{matrix} \leftarrow \text{Top-level} \\ \leftarrow \text{Cluster 1} \\ \leftarrow \text{Cluster 2} \\ \vdots \\ \leftarrow \text{Cluster } k \end{matrix} .$$

Once the time series vector $(\mathbf{y}_t)_{t>0}$ has been built, using some exogenous variables $(\mathbf{x}_t)_{t>0}$, we train n statistical or machine learning models to forecast each of its component. For $i = 1, \dots, n$ and any time step t, we denote by $\hat{f}_i$ the model which outputs the prediction $\widehat{\mathrm{y}}_t^i = \hat{f}_i(\mathbf{x}_t)$ of y_t^i. The forecasts are then gathered into a vector $\widehat{\mathbf{y}}_t$. Finally, we update them using a linear reconciliation method, which is based on the estimation of a matrix $\mathbf{G}$, so that the final forecasts are $\widetilde{\mathbf{y}}_t = \mathbf{S}\mathbf{G}\widehat{\mathbf{y}}_t$. Similarly to Wickramasuriya et al. (2019), we consider linear reconciliations such that the reconciliation matrix equal to

$$\mathbf{G} = \left(\mathbf{S}'\boldsymbol{W}^{-1}\mathbf{S}\right)^{-1}\mathbf{S}'\boldsymbol{W}^{-1}, \tag{1}$$

with $\boldsymbol{W}$, a $n \times n$ definite positive matrix. To sum-up, our three-steps approach, illustrated in Figure 2, relies on the estimations of the summing matrix $\mathbf{S}$ (clustering), the forecasting models $\left(\hat{f}_i\right)_{1\leq i\leq n}$ and $\boldsymbol{W}$, needed to define $\mathbf{G}$ (reconciliation). To validate our procedure, we cut the data into three sub-data sets: the training data set is used for the estimation of the clustering structure and of the models, the calibration data set is for the estimation of the reconciliation matrix $\mathbf{G}$. Finally, we assess the relevance of our approach on a testing test.

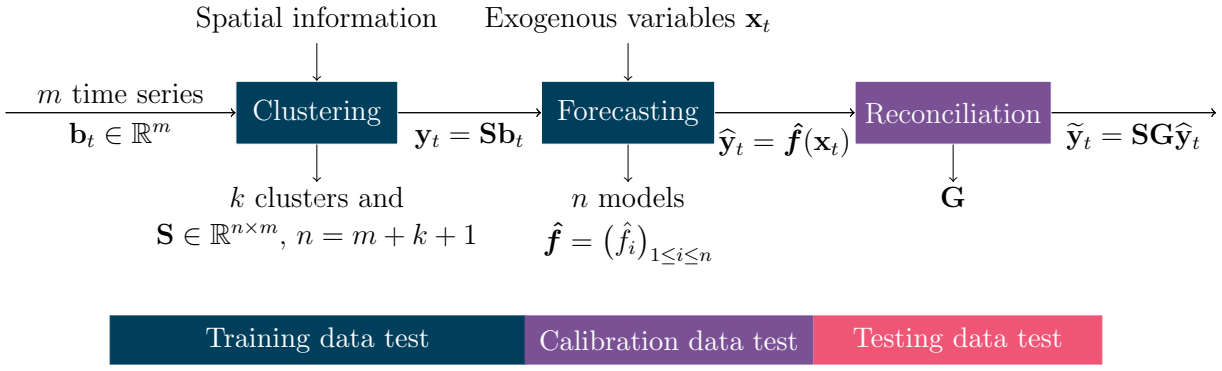


Figure 2: Clustering, Forecasting, and Reconciliation Process Flow.

Hierarchical Spatio-Temporal Clustering of Hourly Data The considered Ward-like hierarchical clustering algorithm (see (Chavent et al., 2018) and Mattera and Franses (2023)) starts with an initial partition with m clusters of singletons. Then, at each step, it aggregates two clusters according to an objective function related to the within-cluster inertia defined as

$$I\left(\mathcal{C}^\ell\right) = (1-\alpha)\,\frac{1}{2|\mathcal{C}^\ell|^2}\sum_{(i,i')\in\mathcal{C}^\ell\times\mathcal{C}^\ell} d_t(i,i')^2 + \;\alpha\frac{1}{2|\mathcal{C}^\ell|^2}\sum_{(i,i')\in\mathcal{C}^\ell\times\mathcal{C}^\ell} d_g(i,i')^2\,, \tag{2}$$

where $d_t(i,i')^2$ and $d_g(i,i')^2$ are squared distances between times series i and i' defined over some temporal and geographical spaces, respectively. The distance $d_g(\cdot,\cdot)$ is the Euclidean distance in the geographical coordinates. For the temporal distance, $d_t(\cdot,\cdot)$, time series b_t^i are split into 24-hour patterns, obtaining 24 multivariate time series $\mathbf{b}_t^i = \left[b_{1,t}^i, b_{2,t}^i, \ldots, b_{24,t}^i\right]^{\mathrm{T}}$ with daily frequency and we define:

$$d_t(i,i') = \frac{1}{24}\sum_{h=1}^{24}\sqrt{2[1-\rho_{i,i'}(h)]}\,, \quad \text{with} \quad \rho_{i,i'}(h) = \mathrm{cor}(b_{h,t}^i, b_{h,t}^{i'})$$

the pairwise correlation coefficient between the series $b_{h,t}^i$ and $b_{h,t}^j$ in the h^{th} hour of the day.

Electricity Demand Forecasting Models Generalized Additive Models are an effective semi-parametric approach for forecasting electricity demand - see, among others, Pierrot and Goude (2011). They model the demand with a sum of (non-linear) variable effects. Meteorological (temperature τ_t^i and smoothed temperature $\bar{\tau}_t^i$) and calendar (day d_t - 1 for Monday, 2 for Tuesday, etc. - and the hour $h_t \in \{1,\ldots,24\}$) variables, gathered in $\mathbf{x}_t$, are used to predict the demand. To model a potential change in the demand after the covid pandemic, we denote by t_0 the time of the beginning of the lockdown (05.01.2020). The more aggregated level the time series, the smoother the demand of electricity and the easier it is to forecast. Therefore, for middle and top-level time series, a greater number of exogenous variables are considered than for bottom-level. The considered models are as follows:

- Bottom-level time series: $\widehat{f}^i(\mathbf{x}_t) = \sum_{d=1}^{7}\sum_{h=1}^{24}\widehat{\alpha}_{d,h}^i\mathbf{1}_{d_t=d}\mathbf{1}_{h_t=h} + \widehat{\alpha}_{\text{covid}}^i\mathbf{1}_{t>t_0}$.
- Middle-level time series: $\widehat{f}^i(\mathbf{x}_t) = \widehat{s}_{\bar{\tau}}^i\left(\bar{\tau}_t^i\right) + \sum_{d=1}^{7}\sum_{h=1}^{24}\widehat{\alpha}_{d,h}^i\mathbf{1}_{d_t=d}\mathbf{1}_{h_t=h} + \widehat{\alpha}_{\text{covid}}^i\mathbf{1}_{t>t_0}$.
- Top-level time series: $\widehat{f}^i(\mathbf{x}_t) = \widehat{s}_{\tau}^i\left(\tau_t^i\right) + \widehat{s}_{\bar{\tau}}^i\left(\bar{\tau}_t^i\right) + \sum_{d=1}^{7}\sum_{h=1}^{24}\widehat{\alpha}_{d,h}^i\mathbf{1}_{d_t=d}\mathbf{1}_{h_t=h} + \widehat{\alpha}_{\text{covid}}^i\mathbf{1}_{t>t_0}$,

with $\widehat{s}_{\tau}^i$ and $\widehat{s}_{\bar{\tau}}^i$ some cubic splines (C^2 piece-wise cubic polynomials). The Penalized Iterative Re-Weighted Least Square (P-IRLS) method of Wood (2017), implemented in the R-package `mgcv`, is used to estimated the models on a training data set.

Optimal reconciliation According to Wickramasuriya et al. (2019), the optimal minimum trace reconciliation is obtained by setting $\boldsymbol{W} = \boldsymbol{\Sigma}$, the variance-covariance matrix of the base forecast errors in Equation (1). Setting $\boldsymbol{W} = \boldsymbol{I}_n$ leads to an orthogonal projection and the Ordinary Least Squares (OLS) reconciliation (Hyndman et al., 2011). Following Figure 2, the estimation of $\boldsymbol{\Sigma}$ is empirical covariance matrix on the calibration data set.

3 Forecasting electricity demand in Alberta, Canada

The data set The Alberta Electric System Operator provides open source electricity demand data[1] for the 42 areas of Alberta at hourly intervals. Temperatures come from the NOAA[2] data base. The training data set contains two full years of data (2019 and 2020); the calibration data set, a year (2021); and the testing data set the remaining data (01.01.2022 - 10.12.2023). Codes and data are provided on a public Github repository[3].

Results Figure 3 shows the cluster structures obtained considering temporal - $\alpha = 0$ in Equation (2) - and spatio-temporal approaches. The accuracy of the unreconcilied forecasts

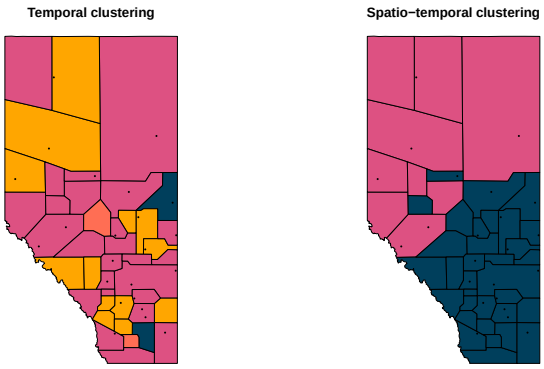


Figure 3: Clustering. Black points correspond the weather stations..

(Base) with respect to different reconciliation approaches – bottom-up (BU), OLS and MinT – using either no cluster structure, temporal cluster structure and spatio-temporal for the top-level series (whole Alberta) are compared in Table 1. For the bottom-up with clusters approach, forecast for the top-level series is obtained from the cluster aggregates. In Table 2, we evaluate for how many areas the use of reconciliation improves the base approach. These results suggest that considerable improvements can be obtained using the MinT and

[1] `https://www.aeso.ca/market/market-and-system-reporting/data-requests/`
[2] National Oceanic and Atmospheric Administration, `https://www.noaa.gov/`
[3] `https://anonymous.4open.science/r/clustering_and_reconciliation-7647`

Clustering	No RMSE · MAPE	Temporal RMSE · MAPE	Spatio-temporal RMSE · MAPE
Base	314 MW · 3.70 %	314 MW · 3.70 %	314 MW · 3.70 %
BU	484 MW · 5.41 %	320 MW · 3.80 %	318 MW · 3.77 %
OLS	311 MW · 3.66 %	312 MW · 3.67 %	311 MW · 3.66 %
MinT	**301 MW · 3.55%**	**302 MW · 3.56 %**	**291 MW · 3.42 %**

Table 1: RMSE and MAPE on testing data set for top-level series (Alberta region).

Clustering	No	Temporal	Spatio-temporal
OLS	54.7 %	54.2 %	54.7 %
MinT	80.1 %	90.4 %	83.3%

Table 2: Percentage of areas for which RMSE has been improved with reconciliation.

spatio-temporal clustering, both for top-level and bottom-level series.

References

Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. (2018). Clustgeo: an r package for hierarchical clustering with spatial constraints. *Computational Statistics*, 33(4):1799–1822.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9):2579–2589.

Mattera, R. and Franses, P. H. (2023). Are african business cycles synchronized? evidence from spatio-temporal modeling. *Economic Modelling*, 128:106485.

Pierrot, A. and Goude, Y. (2011). Short-term electricity load forecasting with generalized additive models. *Proceedings of ISAP power*, 2011.

Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Online learning with signatures for time series forecasting

Nina Drobac[1] & Margaux Brégère[1,2] & Joseph de Vilmarest[3] & Olivier Wintenberger[1]

[1] *LPSM, Sorbonne University, Paris, France* *nina.drobac@sorbonne-universite.fr*
[2] *EDF R&D, Lab Paris-Saclay, France*
[3] *Viking Conseil, Chartres, France*

Résumé. La prévision de consommation d'électricité est essentielle pour assurer l'équilibre sur le réseau électrique. Les évolutions récentes (pandémie de Covid 19, crise énergétique, nouveaux usages électriques) mettent en évidence l'importance de rendre les modèles de prévision adaptatifs. Le cadre de prévision de séries temporelles que nous proposons consiste en un modèle de régression linéaire en une transformation de la série des covariables connue sous le nom de signature. Une signature est une collection d'intégrales itérées, que nous utilisons comme un ensemble de caractéristiques non paramétriques pour les données séquentielles. La signature capture efficacement sa dynamique temporelle et ses modèles d'interaction. Elle a récemment été appliquée en apprentissage automatique en raison de certaines propriétés, notamment son universalité, qui garantit qu'elle peut être combinée avec un modèle linéaire pour approcher n'importe quelle fonction continue du flux de données. Notre approche est basée sur le calcul de la signature sur une fenêtre glissante de largeur fixe, permettant au modèle de s'adapter aux informations récentes. Nous démontrons que les garanties théoriques d'universalité sont préservées. En outre, nous développons un algorithme en ligne qui exploite les propriétés algébriques des signatures pour les calculer plus efficacement. Enfin, nous appliquons notre cadre au problème initial de la prévision de consommation d'électricité.

Mots-clés. Série temporelle, Apprentissage en ligne, Signature, Prévision de la charge.

Abstract. Recent challenges in forecasting electricity load, such as the COVID pandemic and the energy crisis, have highlighted the importance of adaptive predictive models. We propose a time series forecasting framework that involves fitting a linear regression model to a transformation of the covariate series known as the signature. A signature is a collection of iterated integrals that serves as a non-parametric feature set for sequential data, effectively capturing its temporal dynamics and interaction patterns. Recently, it has attracted interest in machine learning applications due to its many useful properties, particularly universality, which guarantees that it can be combined with linear models to approximate any continuous function of the data flow. Our approach is based on computing the signature of a time series

over a sliding window of fixed width, allowing the model to adapt to recent and relevant information. We demonstrate that theoretical guarantees in terms of universality are preserved. Furthermore, we develop an online algorithm that leverages the algebraic properties of signatures to compute them more efficiently. Finally, we will apply our framework to the initial problem of forecasting electricity load.

Keywords. Time series, Signature, Online learning, Load forecasting.

1 Introduction

To balance the supply and demand of electricity and keep the grid operating 24 hours a day, it is crucial to predict the electricity demand as accurately as possible. Sudden changes in customer behavior (e.g. COVID pandemic, energy crisis, energy sobriety) underline a need for adaptive machine learning models for this task.

A common approach to forecast demand is to employ regression models on weather and calendar variables. Since the dependencies being modeled are non-linear, Generalized Additive Models (GAMs) offer a flexible solution by representing the target variable as the sum of smooth functions of covariates; we refer to Wood, 2017. Besides nonlinearity, it is also important to account for the lag in the impact of explanatory variables on the target. For example, due to the thermal inertia of buildings, changes in temperature influence electricity consumption with a delay. One way to account for these past dependencies is to use an exponentially smoothed temperature as a covariate in the GAM, as in Nedellec et al., 2014.

We propose an alternative approach that consists of a simple (possibly regularized) linear regression model on a transformation of covariates called the *signature transform*, which captures both non-linear relationships and past dependencies. Unlike semi-parametric modeling, regression on signatures does not require expert intervention in terms of fixing a basis for GAMs nor choosing smoothing hyper-parameters to model inertia, and is more flexible in modeling interactions.

Signatures have been introduced by Chen, 1958 as a collection of iterated integrals of a path against itself and further developed for rough path theory by Lyons et al., 2007 and others. In short, they are complex mathematical objects that capture key properties of multi-dimensional paths, such as interactions between path components and the order of observed events. Since stochastic processes, and subsequently their (discrete) observations, can be seen as paths, signatures represent a natural feature set for sequential data. For a comprehensive overview of the signature transform in the context of machine learning and statistics, we refer

to Chevyrev and Kormilitzin, 2025.

We introduce a framework for using signatures on sliding windows as a generic feature extraction method for time series forecasting. We begin by presenting signatures and their fundamental properties. Next, we demonstrate that our approach is well theoretically founded by proving that the universal approximation theorem for signatures holds within the sliding window context. Building on this, we leverage the algebraic properties of signatures to develop an efficient online algorithm for their computation. Finally, we will present the use of truncated signatures as feature vectors for time series, combining them with online regression models to forecast electricity consumption.

2 Signatures

We refer to all continuous mappings $x : [s,t] \longrightarrow \mathbb{R}^d$ as continuous paths, and we only consider those of bounded variation (finite length). The *signature* of a path x on $[s,t]$ is defined as the sequence of tensors of increasing order

$$S\left(x_{[s,t]}\right) = \left(1, S^1\left(x_{[s,t]}\right), S^2\left(x_{[s,t]}\right), \ldots, S^k\left(x_{[s,t]}\right), \ldots\right)$$

where, for each integer k,

$$S^k\left(x_{[s,t]}\right) = \int\limits_{s \leq u_1 < \cdots < u_k \leq t} \cdots \int dx_{u_1} \otimes \cdots \otimes dx_{u_k} \in \left(\mathbb{R}^d\right)^{\otimes k} \quad .$$

In other words, the signature of a path is an infinite sequence of iterated integrals of the path against itself, where the collection of k-fold integrals called a *level* is organised as a k-dimensional tensor. It was proven in Lyons, 2014 that the tensor norm of the k-th level of the signature has a factorial decay in k, meaning that for a sufficiently large N, the *truncated signature* $S^{\leq N}\left(x_{[s,t]}\right) := \left(1, S^1\left(x_{[s,t]}\right), \ldots, S^N\left(x_{[s,t]}\right)\right)$ can be used to approximate $S\left(x_{[s,t]}\right)$. Furthermore, thanks to this factorial decay, it can be shown that signatures are elements of a separable Hilbert space of square-summable tensor sequences.

While the definition itself is complex, the truncated signature can be viewed as a transformation of a path that embeds it in a rich (finite-dimensional) space. This results in a set of features that have many useful properties. For instance, Chen's property tells us that when we concatenate two paths $x : [s,t] \longrightarrow \mathbb{R}^d$ and $y : [t,u] \longrightarrow \mathbb{R}^d$, the signature of the resulting path $x * y : [s,u] \longrightarrow \mathbb{R}^d$ is given by $S((x*y)_{[s,u]}) = S\left(x_{[s,t]}\right) \otimes S(y_{[t,u]})$. The time-reversal property states that the inverse of the signature $S(x_{[s,t]})$ is exactly the signature of the path

traversed backwards in time, i.e. $S\left(x_{[s,t]}\right) \otimes S\left(\overleftarrow{x}_{[s,t]}\right) = (1, 0, \ldots)$, where $\overleftarrow{x}(u) = x_{s+t-u}$.

However, the main motivation for using signatures as features for machine learning tasks stems from the following theorem, which says that linear functions on truncated signatures can be used to approximate any continuous function on a compact set of paths.

Theorem 1 (Universal approximation) *Let K be a compact set of bounded variation paths from $[s,t]$ to $\mathbb{R}^d$ such that for every $x \in K$, $x_0 = 0$ and x has at least one monotone coordinate. Let $f : K \to \mathbb{R}$ be continuous. Then, for every $\varepsilon > 0$, there exists a truncation order $N \in \mathbb{N}$ and a vector of coefficients θ, such that*

$$\sup_{x \in K} |f(x) - \langle \theta, S^{\leq N}(x_{[s,t]}) \rangle| \leq \varepsilon \quad .$$

3 Signatures for time series

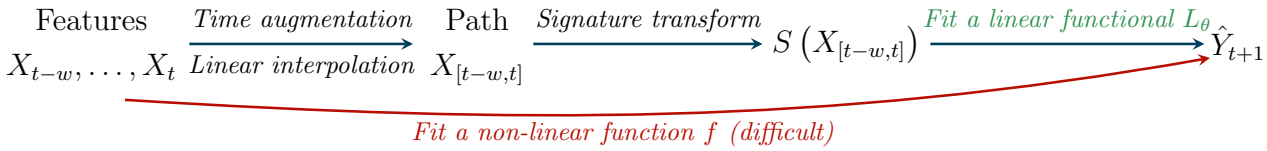
The initial problem of load forecasting can be generalized to the following mathematical setting: we wish to approximate the function f that models the effect of the series of covariates $(X_t)_t$ on the target time series $(Y_t)_t$, i.e. we suppose $Y_t = f(X_{s \leq t}) + \varepsilon_t$. Instead of directly fitting the function f, which is often non-linear and complicated to infer, we take a two step approach. We first use the signature transform to embed the original covariates time series in a more complex space of tensor sequences. The universal approximation theorem now guarantees that this space is rich enough so that f can be approximated by fitting linear functionals on the signature.

In our work, we combine signatures with the concept of *sliding windows* of fixed width w, meaning that at time t, we only consider the last $(w+1)$ observations of the feature time series $(X_{t-w}, \ldots, X_t)$. We chose this approach in order to create a model that is focused on recent and relevant data, that is more adaptive to changes in the time series dynamics and that can be fitted into an online learning setting.

As our observations come at discrete time points, they must first be embedded in continuous time in order to obtain a path. The choice of embedding and path augmentation has been shown to influence the performance of signature-based methods (see Fermanian, 2021). Due to the nature of our application and theoretical guarantees, we opted for linear interpolation of data points, as well as time augmentation on the window. Therefore, when talking about the signature of the d-dimensional time series $(X_t)_t$ over a window $[t-w, t]$, we refer to the signature of the piecewise linear path obtained by interpolating $((0, X_{t-w}), (1, X_{t-w+1}), \ldots,$

(w, X_t)). We denote path as $X_{[t-w,t]}$ and its signature as $S\left(X_{[t-w,t]}\right)$.

Our approach can be summarized as follows:



From a theoretical standpoint, calculating signatures of the feature time series on a sliding window results in a new series $\left(S\left(X_{[t-w,t]}\right)\right)_t$ for which we show the following:

Proposition 1 (Stationarity of signatures) *Let $(X_t)_t$ be a multidimensional discrete time series with strictly stationary increments, then $\left(S\left(X_{[t-w,t]}\right)\right)_t$ is strictly stationary.*

We also demonstrate that fitting a linear regression model on the signature is theoretically well supported because the universality theorem still holds in our setting.

Theorem 2 (Universality on sliding windows) *Let $(X_t)_t$ be a multidimensional bounded time series. Let $U_w = \left\{X_{[t-w,t]} \mid t \in \mathbb{Z}\right\}$ denote the set of paths on all windows and let $f : U_w \longrightarrow \mathbb{R}$ be continuous. For every $\varepsilon > 0$ there exists a linear functional L_θ such that*

$$\sup_t \left| f\left(X_{[t-w,t]} - X_{t-w}\right) - L_\theta\left(S\left(X_{[t-w,t]}\right)\right) \right| \leq \varepsilon \quad .$$

To bridge the gap between theory and application, we note that while the tensor structure of signatures is crucial for deriving theoretical guarantees, we work with the truncated signature as a vector of features. The truncation order N and window size w are hyperparameters of our model.

Moving on to implementation, we develop an online algorithm that exploits the algebraic properties of signatures for more efficient calculation. As we slide the window along our path of covariates, we update the signature in two steps: we overwrite the first segment by using the time reversal property and we then concatenate the new segment to the end of the path using the Chen's identity. We implement the procedure in Python using the `iisignature` library by Reizenstein and Graham, 2018. Additionally, we highlight that the algorithm can be seamlessly integrated with online regression methods, which are widely used in forecasting due to the continuous arrival of data.

Algorithm 1: Signatures on sliding windows

input : signature on the previous window $S = S^{\leq N}\left(X_{[T-w-1,T-1]}\right)$

calculate $S^{\leq N}\left(\overleftarrow{X}_{[T-w-1,T-w]}\right)$;

concatenate $S = S^{\leq N}\left(\overleftarrow{X}_{[T-w-1,T-w]}\right) \otimes S$; `// time reversal`

calculate $S^{\leq N}\left(X_{[T-1,T]}\right)$;

concatenate $S = S \otimes S^{\leq N}\left(X_{[T-1,T]}\right)$; `// Chen`

output: signature on the current window $S = S^{\leq N}\left(X_{[T-w,T]}\right)$

The next phase of our work consists of applying our model to electricity load forecasting.

4 Acknowledgements

This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Sklodowska-Curie Grant Agreement N0 101081674.

References

Chen, K.-T. (1958). Integration of paths—a faithful representation of paths by non-commutative formal power series. *Transactions of the American Mathematical Society*, 89:395–407.

Chevyrev, I. and Kormilitzin, A. (2025). A primer on the signature method in machine learning.

Fermanian, A. (2021). Embedding and learning with signatures. *Computational Statistics Data Analysis*, 157:107148.

Lyons, T. (2014). Rough paths, signatures and the modelling of functions on streams.

Lyons, T., Caruana, M., and Lévy, T. (2007). *Differential Equations Driven by Rough Paths*.

Nedellec, R., Cugliari, J., and Goude, Y. (2014). Gefcom2012: Electric load forecasting and backcasting with semi-parametric models. *International Journal of Forecasting*, 30(2):375–381.

Reizenstein, J. and Graham, B. (2018). The iisignature library: efficient calculation of iterated-integral signatures and log signatures.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman and Hall/CRC, 2 edition.

Prévision à court terme de consommations électriques locales à l'aide de modèles globaux

Roxane Goffinet [1] & Bachir Hamrouche [2] & Julie Keisler [3]& Guillaume Lambert [4]

[1] *EDF - LSCE, France, roxane.goffinet-bourguignon@edf.fr*
[2] *EDF, France, bachir.hamrouche@edf.fr*
[3] *INRIA, France, julie.keisler@inria.fr*
[4] *EDF, France, guillaume.lambert@edf.fr*

Résumé. La prévision de la consommation d'électricité est essentielle pour la gestion du réseau de distribution d'électricité. Cependant, plus le nombre de nœuds du réseau augmente, moins la modélisation individuelle est adaptée. Pour répondre aux défis liés à la prévision d'un nombre croissant de séries temporelles, il y a un intérêt marqué dans l'industrie et la littérature pour la modélisation globale. Cette dernière repose sur l'apprentissage conjoint de l'ensemble des séries temporelles par un unique modèle, dit modèle global, contrairement à la modélisation locale où un modèle est estimé par série temporelle. On parle alors de modèles locaux.

Dans cet article, nous explorons la modélisation globale comme alternative à la modélisation locale dans le cas de la prévision de la consommation d'électricité au niveau des postes sources en France métropolitaine. Ces derniers forment l'interface entre le réseau de transport, géré par RTE (Réseau de transport d'électricité), et le réseau de distribution, géré par ENEDIS. Contrairement aux modèles locaux, les modèles globaux exploitent les schémas partagés entre plusieurs séries temporelles, permettant une approche généralisée de la prévision, avec l'avantage supplémentaire de prévoir des séries qui ont peu d'historique.

Nous testons plusieurs techniques de modélisation globale, comprenant les modèles additifs généralisés (GAM), les forêts aléatoires et les réseaux de neurones, puis nous les comparons à une approche locale basée sur des GAM individuels. Nous évaluons également le nombre de séries temporelles utilisées dans l'apprentissage par rapport aux performances de ces modèles.

Mots-clés. Modèles globaux, Séries temporelles, Prévision de consommation électrique

Abstract. Electricity consumption forecasting is crucial for managing distribution networks. Traditional forecasting models are becoming less practical as the amount of data and the number of nodes requiring forecasts grow exponentially. To address these challenges, there has been a growing interest in global models within the literature.

In this article, we explore the potential of global modelling approaches as a more efficient alternative for forecasting energy loads at the DSO/TSO interface. Unlike local models, global models leverage shared patterns across multiple time series, allowing for a scalable and generalized approach to forecasting, with the added benefit of enabling zero-shot learning. We compare the performance of several modelling techniques in a global setting, including Generalized Additive Models (GAMs), Random Forests and Neural Networks and benchmark them against a local approach based on GAMs.

Our study provides encouraging evidence for adopting global models for electricity load forecasting in large-scale distribution networks, as they present a scalable and efficient solution.

Keywords. Global Models, Time series, Electric load consumption

1 Introduction

La prévision de séries temporelles joue un rôle clef dans de nombreux domaines. Prédire avec précision les valeurs futures d'une variable en fonction de ses valeurs passées et de variables explicatives est essentiel pour une prise de décision efficace. Ces dernières années, l'accès à des données à grande échelle et les avancées en apprentissage automatique ont considérablement élargi le potentiel des modèles de prévision. Traditionnellement, chaque série temporelle est modélisée individuellement et indépendamment selon une approche dite locale. Cette approche est pertinente pour modéliser les spécificités de chaque série, mais le coût computationnel et le stockage associés explosent lorsque le nombre de séries temporelles devient important. A l'inverse, les modèles globaux, qui apprennent à prédire conjointement plusieurs séries temporelles, offrent une alternative prometteuse en exploitant les motifs partagés entre les séries. La modélisation globale permet non seulement de réduire les coûts informatiques, mais aussi de produire des prévisions pour des séries non observées, à condition de choisir des données d'entrée du modèle global accessibles pour ces séries non vues.

Dans le domaine de l'énergie électrique, la prévision d'un grand nombre de séries temporelles est particulièrement importante, car la demande doit être prévue sur l'ensemble du réseau électrique. Cet article considère l'interface entre le réseau de transport et le réseau de distribution de l'électricité en France métropolitaine. Cette dernière a lieu dans des infrastructures appelées postes sources, au nombre d'environ 2300. L'approche par modélisation globale s'avère particulièrement pertinente dans le cas où de nouveaux postes sources sont intégrés au réseau électrique, ou lorsque la prévision s'étend à un niveau plus profond dans le réseau, synonyme d'un nombre encore plus important de séries temporelles à prévoir.

Dans ce travail, nous étudions les méthodes suivantes de prévision de séries temporelles par régression : les Modèles Additifs Généralisés (GAM), les Forêts Aléatoires (RF), et les Réseaux de Neurones (NN). Nous proposons d'étudier deux axes d'analyse sur la modélisation globale de séries temporelles. En premier lieu, nous comparons les modèles ci-dessus appris sur le même ensemble d'entraînement, en distinguant les performances observées sur les séries présentes dans les données d'entraînement et celles observées sur des séries non vues dans l'apprentissage. Dans un second temps, nous examinons l'évolution des performances des modèles globaux selon le nombre de séries utilisées dans la période d'entraînement. L'objectif est d'atteindre une modélisation performante tout en restant frugale dans son estimation.

2 Etat de l'art

Plusieurs études ont démontré l'efficacité des modèles globaux dans la prévision de séries temporelles. Hewamalage et al. [1] ont montré que les modèles globaux surpassent les modèles locaux, en particulier sur des séries qui présentent un historique peu profond ou dont les motifs antérieurs sont peu connus. Ces résultats sont confirmés par Montero-Mans et al. [2], qui comparent des modèles globaux de séries temporelles (comme des modèles Autorégressifs Linéaires ou des Réseaux Profonds Autoregressifs) avec des méthodes locales comme ARIMA et ETS. Les modèles linéaires globaux s'avèrent alors particulièrement effi-

caces, atteignant une précision comparable avec moins de paramètres à estimer, une fois le nombre de paramètres ramené à l'échelle de toutes les séries. L'article souligne notamment que les modèles globaux maintiennent une capacité de généralisation même pour de grands volumes de données et sont donc moins sujets au suraprentissage, contrairement aux modèles locaux. Ils montrent enfin que l'estimation de plusieurs modèles locaux peut être vue comme un simple cas particulier de l'estimation d'un modèle global et qu'en théorie, un modèle global devrait pouvoir être au moins aussi performant qu'un modèle local, à condition qu'il soit suffisamment complexe. D'autres avancées récentes soutiennent également la robustesse des modèles globaux. Par exemple, Cargan et al.[3] montrent que les modèles globaux entraînés sur des données de sources et de natures différentes produisent des prévisions cohérentes et précises, en particulier avec les architectures CNN et LSTM. D'autres utilisent les transformers, comme PatchTST [4].

Dans le domaine spécifique de la prévision de la consommation électrique, les modèles traditionnels comme les Modèles Additifs Généralisés (GAM) ont montré de très bonnes performances, comme souligné par Goude et al. [5] ainsi que par Lambert et al. [6]. Les modèles basés sur les arbres et les réseaux de neurones feed-forward, comme démontré dans des articles tels que ceux réalisés par Lahouar et al. [7] et Kaytez et al.[8], sont également des solutions fiables pour la prévision de la consommation, bien que leur précision dépende de l'échelle et de la disponibilité des données.

3 Méthodologie

3.1 Modèles locaux et modèles globaux

Dans les modèles de prévision locaux, chaque série temporelle est modélisée indépendamment. Supposons que nous ayons $N \in \mathbb{N}^*$ séries temporelles, notées $\{Y^{(i)}\}_{i=1}^{N}$, où chaque série temporelle $Y^{(i)} = \{y_t^{(i)}\}_{t=1}^{T_i} \in \mathbb{R}^{T_i}$ est composée de T_i observations. L'objectif est de modéliser chaque série séparément afin de faire des prévisions.

Nous modélisons chaque série temporelle $Y^{(i)}$ à l'aide de la fonction $f^{(i)}$:

$$\hat{y}_t^{(i)} = f^{(i)}(y_{t-\tau_1}^{(i)}, \ldots, y_{t-\tau_p}^{(i)}, \mathbf{x}_t^{(i)}) \tag{1}$$

où :

- $\hat{y}_t^{(i)}$ est la valeur prédite pour la i-ème série au temps t.
- $\{y_{t-\tau_1}^{(i)}, \ldots, y_{t-\tau_p}^{(i)}\}$ sont les observations passées de la i-ème série.
- $\mathbf{x}_t^{(i)}$ sont les variables exogènes associées à la i-ème série.
- $f^{(i)}$ est un modèle entraîné spécifiquement pour la i-ème série.

Cependant, dans les modèles de prévision globaux, un seul modèle est entraîné sur toutes les séries temporelles simultanément. L'objectif est d'exploiter les informations de toutes les

séries pour effectuer les prévisions. Au lieu d'entraîner N modèles indépendants, une seule fonction globale g est apprise :

$$\hat{y}_t^{(i)} = g(y_{t-\tau_1}^{(i)}, \ldots, y_{t-\tau_p}^{(i)}, \mathbf{x}_t^{(i)}) \tag{2}$$

3.2 Métrique d'évaluation

La précision des prévisions est évaluée à l'aide de l'Erreur Absolue Moyenne Normalisée (NMAE) : NMAE $= \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t| / \bar{y}$, où y_t et $\hat{y}_t$ sont respectivement les valeurs réelles et prédites, et $\bar{y}$ est la valeur moyenne réelle. Cette métrique étant normalisée, elle est intéressante pour comparer des prévisions sur des séries d'odre de grandeur différents. La distribution des scores NMAE (un score par série) est analysée à l'aide des statistiques suivantes : minimum, premier quartile, médiane, troisième quartile, maximum. Pour les réseaux de neuronnes nécessitant une seule métrique pour l'optimisation, c'est le score nmae de la prédiction individuelle qui est utilisé.

4 Résultats

4.1 Dataset

Le jeu de données regroupe 2 241 séries temporelles représentant la consommation électrique au niveau des postes sources. Les emplacements de tous les postes de transformation peuvent être visualisés sur la carte présente en figure 1.

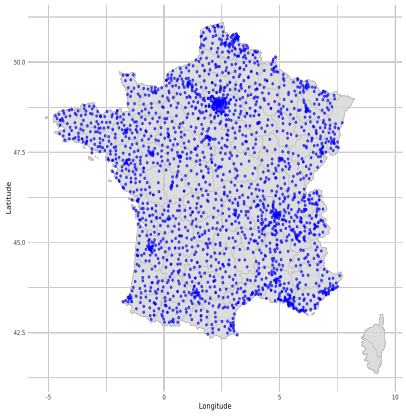


Figure 1: Carte de la localisation des postes de transformation.

Pour chaque série, nous disposons en plus de la consommation d'une trentaine de variables explicatives, incluant : des variables calendaires, des variables météorologiques issues des

réanalyses ERA5 ainsi que des variables de l'historique. Toutes ces variables sont décrites en annexe.

Ces données sont collectées à une fréquence de 30 minutes, soit 48 instants par jour et couvrent une période de six ans, du 1er janvier 2017 au 31 décembre 2023.

Nous effectuons avant l'entraînement et l'évaluation une normalisation par série en utilisant une variante de la formule min-max comme étape de prétraitement. La formule est donnée par : $v_n = \frac{v - 0.9 \times \min_{\text{série}}}{1.1 \times \max_{\text{série}} - 0.9 \times \min_{\text{série}}}$

La normalisation des consommations est d'une importance capitale dans la modélisation globale car elles ont des ordres de grandeur différents. Elle est également nécessaire pour les variables explicatives en étape préliminaire à l'utilisation des réseaux de neurones pour garantir une contribution plus uniforme des variables, prévenir l'instabilité numérique et augmenter la vitesse de convergence du modèles. Il est important de noter qu'une étape de post-traitement est nécessaire par la suite pour convertir la valeur de sortie normalisée en la valeur prédite réelle. Dans le cas d'un poste source nouvellement ajouté au réseau, l'information sur sa consommation minimale et maximale est prévisible par les équipes métiers.

Le jeu de données a été divisé en trois périodes : apprentissage sur les années 2017 à 2021, validation sur l'année 2022, et test sur l'année 2023.

4.2 Comparaison des modèles de prévision globaux

La comparaison des modèles globaux passe par la comparaison des modèles globaux entre eux, mais aussi par la comparaison avec une modélisation locale de référence. La modélisation de référence est l'estimation d'un GAM par série temporelle, soit 2 241 modèles, partageant tous la formule GAM suivante :

$$\begin{aligned} y_t = & \sum_{i=1}^{5}\sum_{j=1}^{11} \alpha_i 1_{TypeJour_t=i, Periode_t=j} + \sum_{i=1}^{2} f_i(PositionDansAnnee_t) 1_{JourOuvre_t=i} + \\ & f_4(Temp_t) + f_5(Temp95_t) + f_6(Temp99_t) + f_7(TempMin_t, TempMax_t) + \\ & + f_8(Vent_t) + Nebulosite_t + f_9(Radiation_t) + \gamma_1 ConsoJ2_t + \gamma_2 ConsoJ7_t + \varepsilon_t, \end{aligned} \tag{3}$$

avec y_t la consommation d'électricité du poste source considéré au temps t. Les effets sont obtenus par projection dans des bases de splines classique (thin plate splines) à l'exception de f_1 et f_2 qui sont issus d'une base de splines cycliques pour tirer profit de la nature cyclique de la variable $PositionDansAnnee_t$ évoluant linéairement de 0 le 1er janvier à 1 le 31 décembre. Le détail des variables explicatives est disponible en annexe.

La modélisation par un modèle GAM global utilise la même formule précédente avec en plus les consommations retardées de 3, 4, 5, 6, 14, et 28 jours. La modélisation globale s'appuie sur des consommations retardées nombreuses et profondes pour se généraliser à des séries non vues. La forêt aléatoire globale est constituée de 100 arbres avec une profondeur maximum fixée à 20 éléments. Pour le réseau de neurones global, l'architecture utilisée est une couche de 30 neurones avec comme fonction d'activation *leakyReLu*. Ces 2 derniers modèles utilisent l'ensemble des variables présentées en annexe.

L'apprentissage est réalisé sur un échantillon de 500 séries temporelles choisies aléatoirement. La prévision est donc effectuée pour 500 séries vues lors de l'estimation des modèles globaux et 1 741 séries non vues. Les performances mesurées par le NMAE sont regroupées dans le tableau 1.

Modèles		NMAE min	NMAE q25	NMAE médiane	NMAE q75	NMAE max
Référence		2.52 %	5.30 %	6.58 %	8.87 %	99.05 %
GAM	Vues	3.28 %	7.61 %	9.33 %	12.42 %	326.43 %
	Non vues	4.23 %	7.46 %	9.16 %	12.70 %	914.54 %
Forêts aléatoires	Vues	2.92 %	6.71 %	9.50 %	14.52 %	70.80 %
	Non vues	3.98 %	7.23 %	10.17 %	14.77 %	88.61 %
Réseaux de neurones	Vues					
	Non vues					

Table 1: Comparaison des performances en NMAE (%) des modèles.

Nos résultats montrent qu'en médiane, les NMAE des modèles globaux ne sont pas particulièrement mauvais. La différence entre la modélisation locale et les modèles globaux, évalués sur des séries utilisées pendant l'entraînement, est inférieure à 3%. Toutefois, il est manifeste que certaines séries sont mal prédites par les GAM globaux. En examinant de plus près, on constate que ce sont les mêmes séries qui présentent les pires erreurs dans tous les modèles, ce qui suggère que certaines séries sont intrinsèquement plus difficiles à prédire que d'autres. En revanche, il est intéressant de noter que les erreurs les plus importantes des forêts aléatoires sont systématiquement inférieures à celles de la référence, qu'elles aient été vues ou non pendant l'entraînement. Néanmoins, en se basant sur le quantile 75%, les GAM globaux semblent offrir de meilleures performances que les forêts aléatoires globales. Ces résultats sont particulièrement prometteurs, surtout lorsqu'on prend en compte les gains en complexité et en temps d'exécution associés aux modèles globaux.

4.3 Influence du nombre de séries lors de l'entraînement des modèles

Une question importante inhérente à la modélisation globale est le nombre de séries temporelles à utiliser lors de l'apprentissage. De plus, du point de vue frugalité informatique, il est également intéressant d'observer s'il y a une limite dans le gain de performances lors de l'ajout de données d'entraînement. Pour y répondre, nous avons entraîné nos modèles en utilisant respectivement 1, 5, 10, 20, 50 et 70 séries, puis évalué leurs performances sur un ensemble de 80 séries non utilisées lors de l'apprentissage. Les résultats sont présentés dans la figure 2. Nous mettons pour comparaison la distribution des erreurs de notre modélisation locale de référence.

On observe que les GAM globaux nécessitent très peu de séries pour offrir de bonnes performances par rapport à la modélisation locale. Dès 10 séries, les erreurs médianes sont comparables à celles des modèles GAM locaux et la distribution des erreurs est moins étalée. Les forêts aléatoires obtiennent également des prévisions satisfaisantes, bien que moins précises que le modèle GAM global. En revanche, les réseaux de neurones exigent un volume de

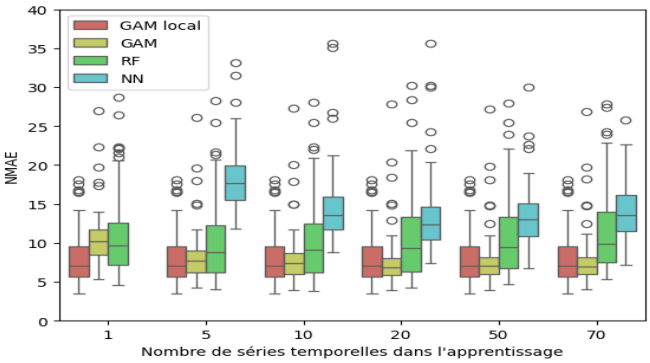


Figure 2: Evolution de la distribution des erreurs NMAE (%) en fonction du nombre de séries utilisées dans l'entraînement.

données beaucoup plus important et présentent des performances médiocres lorsqu'ils sont entraînés sur une seule série (avec un NMAE supérieur à 100%). Enfin, indépendemment du modèle, on observe une stagnation de l'amélioration des prévisions en fonction du nombre de séries dans l'apprentissage.

5 Discussion

Les modèles globaux montrent des performances qui rivalisent parfois avec celles des modèles locaux, en particulier lorsqu'ils utilisent des techniques telles que les GAM ou les forêts aléatoires. Ce résultat est d'autant plus prometteur que les modèles globaux sont beaucoup plus simples à déployer à grande échelle, ce qui les rend particulièrement adaptés à des applications comme l'analyse de la consommation électrique individuelle. Nous avons également constaté qu'il est possible d'obtenir des performances proches de celles des modèles locaux avec seulement 50 séries de données, ce qui positionne ces modèles comme des solutions d'avenir capables de surmonter les limitations liées à l'évolution tout en maintenant des résultats satisfaisants.

Néanmoins, certains modèles globaux montrent de faibles performances pour certaines séries de données, avec des NMAE atteignant parfois des valeurs supérieures à 80%. Cela suggère qu'il serait pertinent d'explorer des approches de partitionnement des données en groupes plus homogènes dans de futures études, afin d'évaluer si cette stratification pourrait améliorer la performance globale. Une telle approche permettrait également de paralléliser l'entraînement, ce qui accélérait davantage le processus.

Par ailleurs, il pourrait être judicieux d'examiner l'intégration d'autres types de modèles, tels que l'apprentissage profond ou l'Apprentissage Automatique Automatisé, qui semblent prometteurs. Ces techniques permettent une optimisation fine de la sélection des caractéristiques, de la conception de l'architecture et des hyperparamètres, facilitant ainsi l'identification de réseaux neuronaux profonds (DNN) plus performants pour la prévision de la charge [8], [9], [10] et [11].

Bibliographie

[1] Hewamalage, H., Bergmeir, C., Bandara, K.: 'Global Models for Time Series Forecasting: A Simulation Study', Pattern Recognition, 2022, (124), pp 108441

[2] Montero-Manso, P., Hyndman, R.: 'Principles and algorithms for forecasting groups of time series: Locality and globality', International Journal of Forecasting, 2021, 6, (37), pp 1632-1653

[3] Cargan, T., Landa-Silva, D., Triguero, I.: 'Local-global methods for generalised solar irradiance forecasting', Applied Intelligence, 2024, 2, (54), pp 1-23

[4] Nie, Y., Nguyen, N., Sinthong, P., and al: 'A Time Series is Worth 64 Words: Long-term Forecasting with Transformers', arXiv abs/2211.14730, 2022

[5] Goude, Y., Nedellec, R. & Kong, N. Local short and middle term electricity load forecasting with semi-parametric additive models. IEEE Trans. Smart Grid 5, 440–446.

[6] Lambert, G., Hamrouche, B. & de Vilmarest, J. Frugal day-ahead forecasting of multiple local electricity loads by aggregating adaptive models. Sci Rep 13, 15784 (2023).

[7] Lahouar, A., Ben Hadj Slama, J.: 'Day-ahead load forecast using random forest and expert input selection', Energy Conversion and Management, 2015, (103), pp 1040-1051

[8] Kaytez, F., Cengiz Taplamacioglu, M., Cam, E., and al: 'Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines', International Journal of Electrical Power and Energy Systems, 2015, (67), pp 431-438

[9] Keisler, J., Claudel, S., Cabriel , G., and al: 'Automated Deep Learning for Load Forecasting', arXiv abs/2405.08842, 2024

[10] Haifeng, C., Chollet, F.: 'AutoKeras: an AutoML library for deep learning', J. Mach. Learn. Res., 2023, 1,(24), pp 1532-1537

[11] Erickson, N., Mueller, J., Shirkov, A., and al: 'AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data', arXiv abs/2003.06505, 2020

[12] Zimmer, L., Lindauer, M., Hutter, F.: 'Auto-Pytorch: Multi-Fidelity MetaLearning for Efficient and Robust AutoDL', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, (43), pp 3079-3090

[13] Chiang, A.: 'Generalized Additive Models: An Introduction With R', Technometrics, 2012, (49), pp 360-361

[14] Breiman, L.: 'Random Forests', Applied Intelligence, 2001, (45), pp 5-32

6 Annexe

Présentation des variables

- $TypeJour_t$: variable catégorielle représentant le jour de la semaine : lundi, mardi à jeudi, vendredi, samedi, et dimanche.
- $Periode_t$: variable catégorielle représentant 11 périodes de l'année : heure d'hiver, heure d'été avant les vacances scolaires d'Août, les 5 semaines d'Août, heure d'été après les vacances scolaires d'Août, vacances de Noël avant Noël, entre Noël et nouvel an, et après nouvel an.
- $PositionDansAnnee_t$: variable numérique évoluant linéairement de 0 le 1er janvier à 1 le 31 décembre.
- $JourOuvre_t$: variable catégorielle indiquant si le jour est ouvré ou non.
- $Temp_t$: variable numérique de la température mesurée à 10 mètres du sol.
- $Temp95_t$, $Temp99_t$: variables numériques de la température $Temp_t$ lissée exponentiellement de coefficient $\alpha = 0.95$ et 0.99 respectivement, c'est-à-dire $Temp95_t = \alpha Temp95_{t-1} + (1-\alpha)Temp_t$.
- $TempMin_t$, $TempMax_t$: variables numériques de la température minimale et maximale journalière.
- $Vent_t$: variable numérique de la norme de la vitesse du vent à 10 mètres en m/s.
- $Nebulosite_t$: variable numérique de la nébulosité en octa.
- $Radiation_t$: variable numérique de la radiation solaire en Jm^{-2}.
- $ConsoJX_t$: variable numérique de la consommation retardée de X jours.
- $longitude_t$, $latitude_t$: coordonnées géographiques du poste source.
- $PeriodeEteHiver_t$: variable catégorielle de 2 périodes dans l'année : heure d'été et heure d'hiver.

Rethinking Synthetic Twins: Average Treatment Effect Estimation with Latent Representations Learning

Ouali Maha [1] , Ghattas Badih [2] , Flachaire Emmanuel [3] , Bozzi Laurent [4] , Charpentier Philippe [5] ,

[1] *EDF R&D - Aix Marseille School of Economics , maha.ouali@edf.fr*
[2] *Aix Marseille School of Economics, badihghattas@gmail.com*
[3] *Aix Marseille School of Economics, emmanuel.flachaire@gmail.com*
[4] *EDF R&D, laurent.bozzi@edf.fr*
[4] *EDF R&D, philippe.charpentier@edf.fr*

Résumé. La demande croissante de flexibilité énergétique a conduit au développement de programmes de réduction de la consommation en période de pointe. EDF R&D a lancé plusieurs initiatives utilisant les données des compteurs Linky. Cependant, évaluer leur impact reste difficile en raison des modèles de consommation individualisés, du biais de sélection dû à la participation volontaire et des limites des méthodes classiques d'inférence causale. Nous simulons un environnement contrôlé pour examiner ces défis et comparons les approches traditionnelles avec des méthodes d'apprentissage automatique comme SyncTwin, qui crée des jumeaux synthétiques pour gérer les variables cachées. Nos résultats montrent que les séries temporelles complexes défient les méthodes classiques, surtout avec des variables non observées influençant le traitement. Nous expliquons théoriquement pourquoi ces approches échouent, en raison de covariables inaccessibles influençant le score de propension. Nos travaux ouvrent la voie à de nouvelles techniques pour améliorer l'évaluation des programmes d'économie d'énergie.

Mots-clés. Inférence causale , Contrôle synthétique, Intelligence Artificielle

Abstract. The increasing need for energy flexibility has led to demand-response programs aimed at reducing peak consumption. EDF R&D has implemented initiatives using Linky smart meter data. However, assessing their impact is challenging due to individualized consumption patterns, selection bias from voluntary participation, and limitations of conventional causal inference methods. We simulate a controlled environment to expose these challenges and compare traditional methods with machine learning approaches like SyncTwin, which creates synthetic twins to manage hidden confounders. Our results show that complex time series challenge classical methods, especially with unobserved confounders influencing treatment. We provide a theoretical explanation for their failure due to unmeasured covariates affecting the propensity score. Our work paves the way for advanced techniques to improve energy-saving program evaluations.

Keywords. Causal inference, synthetic control, Artificial intelligence

1 Introduction

1.1 Causal inference setting

Causal inference aims to estimate the effect of an intervention when treatment is not randomly assigned, making direct comparisons between treated and control groups potentially biased. We consider an observational study with $N+1$ individuals indexed by $i \in \{1, \ldots, N+1\}$. For simplicity, we assume a single treated individual, with the rest forming the control group. Let T_i denote the treatment indicator, where $T_i = 1$ if individual i is treated and $T_i = 0$ otherwise. All treatments are assumed to be initiated at time T_0.

The potential outcomes framework defines Y_i^1 as the outcome if individual i were treated and Y_i^0 as the outcome if untreated. The observed outcome is then given by:

$$Y_i = \mathbf{1}_{T_i=1} Y_i^1 + \mathbf{1}_{T_i=0} Y_i^0.$$

We assume that observable variables X and latent variables W jointly influence both treatment assignment and outcomes. For instance, individuals with strong environmental awareness may opt into the treatment, introducing selection bias.Because W is unobserved, we cannot directly adjust for it to satisfy the backdoor criterion (Pearl, 2009), as is done in propensity score methods (Rosenbaum & Rubin, 1983).

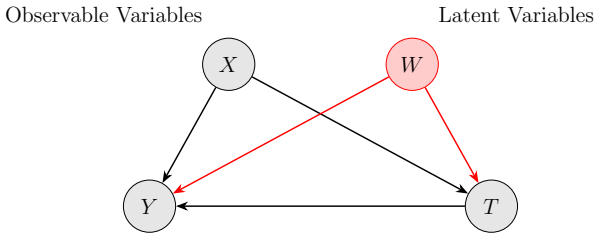


Figure 1: Directed Acyclic Graph (DAG) representing the causal structure in our work context. Observable variables (X) and latent variables (W) influence both the treatment allocation (T) and the outcome (Y).

1.2 Synthetic Control Method (SCM)

Synthetic control methods (Abadie, Diamond & Hainmueller, 2010) estimate the counterfactual outcome of a treated unit by constructing a weighted combination of control units. These weights are derived from pre-treatment observational data,primarily the outcomes of interest Y, and optionally, other covariates X, to best approximate the treated unit prior to intervention.
In our specific case study, the primary source of information comes from high-frequency electricity consumption load curves. These time series data encode rich behavioral patterns and are therefore the most informative covariates available. Accordingly, we define X to consist of the pre-treatment values of the outcome variable Y, i.e., $X = (Y_t)_{t<T_0}$. While additional

covariates could be incorporated, our focus remains on the pre-treatment load curves due to their relevance and predictive power in estimating treatment effects.

We consider a population of $N+1$ units. Let $Y_{1,t}$ denote the outcome of the treated unit at time t, and $Y_{j,t}$ the outcomes of control units $j \in \{2, \ldots, N+1\}$. In this setting, the covariates X correspond to the vector of pre-treatment outcomes, i.e., Y_t for $t < T_0$, where T_0 is the time of intervention.

The synthetic control weights $(b_j)_{j=2}^{N+1}$ are chosen such that:

$$Y_{1,t} \approx \sum_{j=2}^{N+1} b_j Y_{j,t}, \quad \text{for } t \in \{0, \ldots, T_0\}. \tag{1}$$

To ensure interpretability and stability, these weights are typically constrained to form a convex combination:

$$b_j \geq 0, \quad \sum_{j=2}^{N+1} b_j = 1. \tag{2}$$

This convexity constraint may be relaxed by introducing regularization techniques such as Lasso or Elastic Net (Doudchenko & Imbens, 2016), or other penalization approaches (Abadie & L'Hour, 2021), to improve robustness and encourage sparsity. While regularization may yield better performance in high-dimensional settings, it can also introduce instability.

The average treatment effect on the treated (ATT) is defined as:

$$\tau = \mathbf{E}(Y^1 - Y^0 \mid T = 1).$$

Since $\mathbf{E}(Y^1 \mid T = 1) = \mathbf{E}(Y \mid T = 1)$, estimation reduces to recovering the counterfactual mean $\mathbf{E}(Y^0 \mid T = 1)$. SCM relies on the assumption that:

$$\mathbf{E}(Y^0 \mid T = 1, X) = \mathbf{E}\left(\sum_{j=2}^{N+1} b_j Y_j \mid T = 0, X\right),$$

which, using Equation 1, implies:

$$\mathbf{E}(Y^0 \mid T = 1, X) = \mathbf{E}(Y^0 \mid T = 0, X) = \mathbf{E}\left(\sum_{j=2}^{N+1} b_j Y_j \mid T = 0, X\right).$$

Again, X includes all observed covariates, in our case, it consists of pre-treatment values of the outcome variable Y_t for $t < T_0$.

This assumption effectively states that $Y^0, Y^1 \perp T \mid X$—i.e., conditional on X, treatment assignment is assumed to be as good as random. This requires that treated and control units are sufficiently comparable based on their pre-treatment outcomes. Violation of this

assumption, particularly in the presence of unobserved confounders affecting both treatment and outcomes, may result in biased estimates.

1.3 Example

For each individual i, we consider that the outcome of interest (in the absence of treatment) is defined as an AR(1) process, W is our hidden confounder that affects the outcome of interest $W \sim \mathcal{U}[0,1]$ and $\epsilon \sim \mathcal{N}(0,\sigma)$ is a white noise:

$$Y_{i,t}^0 = \begin{cases} \rho Y_{i,t-1}^0 + \epsilon_{i,t}, & \text{if } W_i <= 0.5, \\ \rho Y_{i,t-1}^0 + \alpha t + \epsilon_{i,t}, & \text{otherwise} \end{cases}$$

Treatment starts at time $t = T_0$, T is the treatment indicator, and the treatment allocation is determined by the hidden factor W creating bias between the treated and control group :

$$p(T = 1|W = w) = g(w) = \frac{1}{1 + \exp(-5(w - 0.5))}.$$

1.3.1 Treatment effect

The treated outcome, Y_t^1, is defined as:

$$Y_t^1 = Y_t^0 + \tau \, \mathbf{1}\{T = 1, t >= T_0\},$$

To estimate τ, we compute $\mathbf{E}(Y^0|T = 1)$ theoretically (see Appendix).

$$\mathbf{E}(Y^0|T = 1) = \frac{\mathbf{P}(T = 1|W > 0.5)P(W > 0.5)}{\mathbf{P}(T = 1)}\mathbf{E}(P^2)$$

Now, assume we have a set of positive weights B^* such that (1) holds.

$$\mathbf{E}(\sum_{j=2}^{N+1} b_j Y_j|T = 0) = \sum_{j=2}^{N+1} b_j \frac{\mathbf{P}(T = 0|W > 0.5)P(W > 0.5)}{\mathbf{P}(T = 0)}\mathbf{E}(P^2)$$

Thus, if $\mathbf{E}(P^2) * \mathbf{P}(W > 0.5) \neq 0$:

$$\mathbf{E}(Y^0|T = 1) = \mathbf{E}(\sum_{j=2}^{N+1} b_j * Y_j^0|T = 0) \iff \frac{\mathbf{P}(T = 1|W > 0.5)}{\mathbf{P}(T = 1)} = \sum_{j=2}^{N+1} b_j \frac{\mathbf{P}(T = 0|W_j > 0.5)}{\mathbf{P}(T = 0)} \quad (3)$$

In essence, the optimal weights must balance the propensity scores of the treated and control groups. This means that individuals with a higher chance of receiving the treatment

driven by the hidden confounder W should be assigned larger weights, while those with a lower propensity should receive smaller weights.

However, in the special case where $T \perp\!\!\!\perp W$ (no selection bias),we get the standard synthetic control condition (2) :

$$b_j >= 0 \quad \forall j \quad \text{and} \quad \sum_{j=2}^{N+1} b_j = 1$$

2 General formulation

We denote by W the hidden confounder that affect both the treatment and the outcome of interest. For instance , we suppose that W is continous and $W \sim \mathcal{P}_W$ with a probability density function P. T is a binary treatment viriable and $T|(W = w) \sim \mathcal{B}\text{ernouilli}(g(w))$. $Y^0(w)$ is the potenial outcome in the absence of treatment , under the hidden confounder $W = w$. $Y^0(w) \sim \mathcal{P}_{Y(w)}$.

Independance assumption
the potential outcomes are independant of the treatment assignment conditional on the hidden confounder W :

$$Y^1(w), Y^0(w) \quad \perp\!\!\!\perp \quad T|W = w$$

Consider $N + 1$ units , where $i = 1$ is the treated unit, and $j = 2, \ldots, N + 1$ are the control units. Suppose we have a vector of positive weights $B^* = (b_j)_{j=2,\ldots,N+1}$, such that :

$$Y_{1,t} = \sum_{j=2}^{N+1} b_j Y_{j,t} \text{ for } t \in \{0, \ldots, T_0\}$$

We can show that using this equation, we have the following result :

$$\mathbf{E}(Y^0|T = 1) = \mathbf{E}(\sum_{j=2}^{N+1} b_j Y_j^0|T = 0)$$
$$\Longleftrightarrow$$

$$\int (\frac{g(w)}{\mathbf{P}(T = 1)} - \sum_{j=2}^{N+1} b_j \frac{(1 - g(w))}{\mathbf{P}(T = 0)})\mathbf{P}(W = w)\mathbf{E}(Y^0(w))dw = 0. \tag{4}$$

Moreover, if $T \perp\!\!\!\perp W$ and $\mathbf{E}(Y^0) \neq 0$ then (4) simplifies to : $B^* = (b_j)_{j=2,\ldots,N+1}$, such that :

$$Y_{1,t} = \sum_{j=2}^{N+1} b_j * Y_{j,t} \text{ for } t \in \{0, \ldots, T_0\} \quad \text{and} \quad \textstyle\sum_{j=2}^{N+1} b_j = 1; \tag{5}$$

The above condition (4) ensures that the hidden confounder's influence on treatment assignment is balanced between the treated unit and the synthetic control. If this condition holds, the synthetic control group can reliably estimate the treated unit's counterfactual outcome.

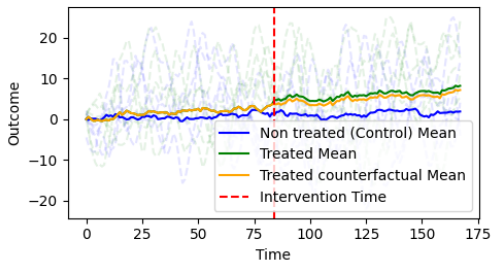
In the absence of confounding—i.e., when $T \perp W$—the condition simplifies to the standard constraint used in traditional synthetic control methods (e.g., Abadie and Diamond)(2). However, in the presence of hidden confounding, the weights must compensate for differences in the distribution of W between treated and control groups using the propensity score ($g(w)$).

3 Benchmark on simulated data :

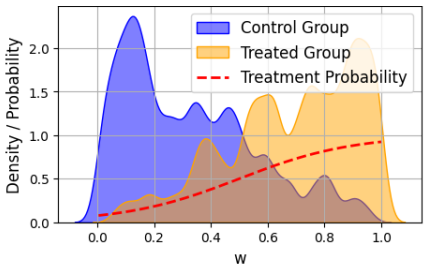
3.1 Process description

We conduct a series of Monte Carlo experiments based on the described simulation setup, repeating the process over 1,000 iterations. In each iteration, we apply the benchmark methods and estimate treatment effects in order to assess their accuracy and robustness. Among the deep learning-based approaches, we include SyncTwin (Qian, Zhang, Bica, Wood, & van der Schaar, 2021), a neural network model designed for estimating individualized treatment effects from longitudinal time-series data. SyncTwin operates in two main phases. In the training phase, it employs an encoder-decoder architecture to learn latent representations of individuals from their pre-treatment time series, capturing temporal dynamics and individual-specific patterns. In the inference phase, these latent embeddings are used to compute similarity-based weights among control units, enabling the construction of personalized synthetic controls. Notably, SyncTwin does not require a pre-training step—the encoder and decoder are trained jointly on the full dataset. For our experiments, the neural network was trained on a simulated dataset of 10,000 individuals. The training process consisted of 20 runs, each executed for 50,000 epochs, and the inference was repeated 1000 time to compute the ATE.

3.2 Graph



(a) Simulated data of an intervention with hidden confounder W



(b) Distribution of W and Treatment Probability

Figure 2: Example of the simulated dataset: $\rho = 0.8, \alpha = 0.05, \sigma = 5, \tau = 0.77$.

3.3 Results:

	ATE	Average time(s)	ATE std	CI	real ATE
did	2.90	0.01	0.36	[2.23 3.63]	0.77
ols	2.90	0.06	0.36	[2.23 3.63]	0.77
synthetic control	3.76	1.85	0.29	[3.22 4.32]	0.77
lasso	3.90	0.87	0.33	[3.30 4.59]	0.77
ridge	3.86	13.43	0.27	[3.37 4.44]	0.77
elastic net	3.90	0.83	0.33	[3.30 4.58]	0.77
synthetic did	3.31	0.67	0.36	[2.64 4.03]	0.77
synthetic dtw	3.45	170.31	0.61	[2.23 4.66]	0.77
Synctwin**	3.92	4233 (training)	0.34	[3.25 4.58]	0.77

By looking the the estimated ATE and the confidence intervals, we see that all the tested methods fail to estimate the treatment effect, even Synctwin which is designed for longitudinal data with hidden confounders, and based on latent representations learning. The methods tend to predict a treatment effect value close to 3, which roughly reflects the average observed difference between the treated and control groups. This bias is due to the presence of an unobserved confounder W, which influences both treatment assignment and the outcome Y (see figure(2)). Since W is not observed, the methods cannot account for its impact, especially in the post-treatment period, where its effect on outcomes is stronger. As a result, the constructed counterfactuals do not accurately reflect what would have happened in the absence of treatment, leading to an overestimation of the true treatment effect.

4 Conclusion

In this study, we investigated the challenges of estimating treatment effects in high-dimensional time-series data, focusing on the evaluation of energy consumption interventions. Through simulated experiments, we demonstrated the limitations of traditional causal inference methods, which struggle to handle hidden confounders and the inherent complexity of individualized load curves.

When testing SyncTwin, the model fell short in accurately predicting counterfactual outcomes. This highlights the difficulty of addressing bias introduced by unobserved confounders, even with advanced machine learning models.

In future work, we will dive deeper into the condition we found on the balancing weights and their relation to the propensity score estimation. Specifically, we will propose an initial model that accounts for differences in propensity scores, offering a more robust solution to mitigate biases caused by unobserved factors.

References

Doudchenko, N., & Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis (No. w22791). National Bureau of Economic Research.

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American statistical Association, 105(490), 493-505.

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. Journal of economic literature, 59(2), 391-425.

Abadie, A., & L'hour, J. (2021). A penalized synthetic control estimator for disaggregated data. Journal of the American Statistical Association, 116(536), 1817-1834.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

Tesei, G., Giampanis, S., Shi, J., & Norgeot, B. (2023). Learning end-to-end patient representations through self-supervised covariate balancing for causal treatment effect estimation. Journal of Biomedical Informatics, 140, 104339.

Qian, Z., Zhang, Y., Bica, I., Wood, A., & van der Schaar, (2021). Synctwin: Treatment effect estimation with longitudinal outcomes. Advances in Neural Information Processing Systems, 34, 3178-3190.

Shi, C., Blei, D., & Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. Advances in neural information processing systems, 32. Pearl, J. (2009). Causality. Cambridge university press.

Chowdhury, R. R., Zhang, X., Shang, J., Gupta, R. K., & Hong, D. (2022, August). Tarnet: Task-aware reconstruction for time-series transformer. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 212-220).

Appendix

$$
\begin{aligned}
\mathbf{E}(Y^0|T=1) = \quad & \mathbf{E}\left(\mathbf{1}_{W\leq 0.5}\cdot P^1 + \mathbf{1}_{W>0.5}\cdot P^2 \mid T=1\right) \\
\overset{P^1,\,P^2\,\perp\, T}{=} \quad & \mathbf{P}(W\leq 0.5|T=1)\cdot \mathbf{E}\left(P^1\right) + \mathbf{P}(W>0.5|T=1)\cdot \mathbf{E}\left(P^2\right) \\
\overset{\mathbf{E}(P^1)=0}{=} \quad & \mathbf{P}(W>0.5|T=1)\cdot \mathbf{E}\left(P^2\right) \\
= \quad & \frac{\mathbf{P}(T=1|W>0.5)P(W>0.5)}{\mathbf{P}(T=1)}\mathbf{E}(P^2)
\end{aligned}
$$

And for the counterfactual approximation :

$$
\begin{aligned}
\mathbf{E}(\sum_{j=2}^{N+1} b_j Y_j|T=0) = \quad & \sum_{j=2}^{N+1} b_j \mathbf{E}(Y_j|T=0) \\
= \quad & \sum_{j=2}^{N+1} b_j \mathbf{E}(\mathbf{1}_{W_j\leq 0.5} * P_j^1 + \mathbf{1}_{W_j>0.5} P_j^2 \mid T=0) \\
\overset{P^1,\,P^2\,\perp\, T}{=} \quad & \sum_{j=2}^{N+1} b_j \mathbf{P}(W_j\leq 0.5|T_j=0)\cdot \mathbf{E}\left(P_j^1\right) + \mathbf{P}(W_j>0.5|T_j=0)\cdot \mathbf{E}\left(P_j^2\right) \\
\overset{\mathbf{E}(P^1)=0}{=} \quad & \sum_{j=2}^{N+1} b_j * \mathbf{P}(W_j>0.5|T_j=0)\cdot \mathbf{E}\left(P_j^2\right) \\
= \quad & \sum_{j=2}^{N+1} b_j \frac{\mathbf{P}(T=0|W>0.5)P(W>0.5)}{\mathbf{P}(T=0)}\mathbf{E}(P^2).
\end{aligned}
$$

Apprentissage fédéré et autres stratégies

Federated Causal Inference: ATE estimation with decentralized observational data

Rémi Khellaf*[1], Julie Josse[1], and Aurélien Bellet[1]

[1]Inria Premedical – Institut National de Recherche en Informatique et en Automatique, Institut National de la Santé et de la Recherche Médicale - INSERM, Université de Montpellier et MBS – France

Résumé

Federated Causal Inference is an emerging paradigm that enables causal effect estimation across decentralized observational datasets without requiring direct access to pooled data. Traditional causal inference methods typically rely on centralized datasets, where all relevant information is aggregated into a single repository. However, in many real-world scenarios, such centralization is not feasible due to privacy regulations, ethical concerns, institutional policies, or the distributed nature of data collection across multiple entities. These constraints necessitate the development of alternative frameworks that can facilitate robust and reliable causal effect estimation while preserving data sovereignty and minimizing the risk of data breaches.
Federated learning (FL) offers a promising solution to these challenges by allowing multiple parties to collaboratively estimate causal effects without directly sharing raw data. Instead of transmitting individual-level data, FL methods involve the exchange of aggregated model parameters or summary statistics, ensuring that sensitive information remains localized while still enabling statistical inference across multiple data silos. This paradigm is particularly relevant in domains such as healthcare, economics, and public policy, where data is often distributed across institutions with stringent privacy requirements.

In this paper, we explore methodologies for performing causal inference in a federated setting, focusing specifically on the estimation of the Average Treatment Effect (ATE). We examine the implementation of Inverse Probability Weighting (IPW) and Doubly Robust Estimators (AIPW) in scenarios where data is siloed across multiple independent data centers. IPW relies on estimated propensity scores to adjust for confounding, while AIPW combines propensity score weighting with outcome modeling to achieve greater robustness and efficiency. These methods are well-established in traditional causal inference but require adaptation to the federated setting, where data heterogeneity and limited information sharing pose additional methodological and computational challenges.
A key challenge in federated ATE estimation is the presence of heterogeneity across participating institutions. Differences in data distributions, treatment assignment mechanisms, or underlying population characteristics can introduce biases that affect the validity of causal estimates. Additionally, the presence of center effects-systematic variations in outcomes due to institutional differences-can further complicate estimation. To address these issues, we assume a common Conditional Average Treatment Effect (CATE) across sites, allowing for site-specific variations in covariate distributions while maintaining a shared treatment effect model.

*Intervenant

Mots-Clés: Apprentissage fédéré, Inférence causale, régression, modèle linéaire, méta, analyses, études observationnelles, score de propension

Analyse Raffinée de Federated Averaging et Extrapolation de Richardson-Romberg Fédérée

Paul Mangold*[1], Alain Durmus[1], Aymeric Dieuleveut[1], Sergey Samsonov[2], and Eric Moulines[1,3]

[1]Centre de Mathématiques Appliquées de l'Ecole polytechnique – Institut National de Recherche en Informatique et en Automatique, Ecole Polytechnique, Centre National de la Recherche Scientifique – France

[2]HSE University – Russie

[3]Mohamed bin Zayed University of Artificial Intelligence – Émirats arabes unis

Résumé

Dans cette communication, nous présentons une nouvelle analyse de avec un pas constant, en nous appuyant sur la propriété de Markov du processus sous-jacent. Nous montrons que les itérés globaux de l'algorithme convergent vers une distribution stationnaire et analysons le biais et la variance qui en résultent. Nous donnons ensuite un développement asymptotique du biais au premier ordre, aussi bien quand les clients sont homogènes qu'hétérogènes. Notre analyse montre que le biais de Federated Averaging, se décompose en deux composantes distinctes : l'une dépendant uniquement du bruit introduit par le gradient stochastique, et l'autre de l'hétérogénéité des clients. Enfin, nous introduisons un nouvel algorithme basé sur la méthode d'extrapolation de Richardson-Romberg afin de diminuer ce biais.

Mots-Clés: Apprentissage fédéré, Optimisation

*Intervenant

Méthodes d'apprentissage faiblement supervisées pour la classification de patients à partir de données de cytométrie en flux

Pierre-André Mikem [1,2] & Christine Keribin [2] & Pascal Massart [2]

[1] *Metafora biosystems, Pépinière Paris Biotech Santé, France*
[2] *Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, France*

Résumé. La cytométrie en flux permet d'analyser des cellules pour détecter des anomalies associées à certaines maladies, mais l'absence d'annotations précises complique l'apprentissage supervisé. Nous proposons une approche faiblement supervisée basée sur le Multiple Instance Learning (MIL) pour classifier les patients selon leur état de santé. Dans ce cadre, chaque patient est représenté comme un bag contenant des cellules (instances), avec en général une annotation disponible uniquement au niveau du patient. Nous proposons deux méthodes principales. Dans une première, nous modélisons la distribution des cellules d'un patient avec un mélange de deux gaussiennes : une pour les cellules saines et une pour les cellules tumorales. L'algorithme EM est utilisé pour déterminer l'annotation des cellules. Cependant, cette approche reste limitée par l'hypothèse de distributions simples. Afin de modéliser les distributions de façon plus flexible, nous proposons dans une seconde approche une méthode combinant auto-encodeur et mélange d'experts. Un auto-encodeur est utilisé pour projeter les cellules dans un espace latent réduit. Un modèle de mélange gaussien regroupe ensuite les cellules en sous-populations, chacune étant associée à un expert. Enfin, la classification d'un patient repose sur la moyenne pondérée des prédictions des experts appliquées à ses cellules. Nous montrerons l'apport de la représentation avec auto-encodeur pour traiter ces données et la comparerons avec d'autres méthodes de référence en MIL que nous appliquerons au cas de la cytométrie.

Mots-clés. Cytométrie en flux, Multiple Instance Learning, Auto-encodeur, Mélange d'experts

Abstract. Flow cytometry allows the analysis of cells to detect abnormalities associated with some diseases, but the lack of precise annotations complicates the learning. We propose a weakly supervised approach based on Multiple Instance Learning (MIL) to classify patients according to their health condition. Each patient is represented in this framework as a bag containing cells (instances), with annotations generally available at the patient level only. We propose two main methods. The first one models the patient's cell distribution with a mixture of two Gaussians: one for healthy cells and one for tumor cells. The EM algorithm is used to determine the annotation of the cells. However, this approach remains limited by the assumption of simple distributions. To model more flexible distributions, we propose a second approach that combines an autoencoder with a mixture of experts. The autoencoder is used to project cells into a reduced latent space. A Gaussian mixture model then regroups the cells into sub-populations, each associated with an expert. Finally, the patient classification is based on the weighted average of the experts' predictions on the cells. We will show the

benefits of using an autoencoder to represent these data and compare our approach with other benchmark MIL methods applied to flow cytometry.

Keywords. Flow cytometry, Multiple Instance Learning, Autoencoder, Mixture of experts

1 Introduction

Certaines maladies telles que les cancers sont liées à des anomalies présentes dans le métabolisme des cellules. Une technique utilisée pour quantifier le métabolisme des cellules est appelée la cytométrie en flux. Elle consiste à faire passer les cellules à travers un laser, où leurs caractéristiques physiques (taille, granularité) et chimiques (marqueurs fluorescents) sont mesurées en temps réel. Les cellules sont ainsi caractérisées sur un ensemble de variables permettant d'identifier les différentes populations et sous-populations présentes dans un échantillon.

Dans ce projet, notre objectif est de développer des méthodes d'apprentissage automatique afin de discriminer les patients selon le stade de gravité de leur maladie, tout en permettant d'extraire des schémas locaux (des cellules) représentatifs de cette gravité. Les données de cytométrie dans ce contexte posent deux types de problème: d'une part l'hétérogénéité des cellules tumorales, d'autre part la fiabilité de leurs annotations. Pour les résoudre, nous proposons une approche par *Multiple Instance Learning*.

Le Multiple Instance Learning (MIL) [Carbonneau et al., 2018] est une méthode d'apprentissage supervisé conçue pour résoudre des problèmes où les données sont regroupées en *bags* (ensembles d'instances) avec un label global, mais sans information sur les labels individuels des instances. Le but du MIL est d'apprendre une fonction qui prédit correctement les labels des bags. Contrairement à l'apprentissage supervisé classique, où chaque instance est directement associée à un label, le MIL repose sur des hypothèses spécifiques pour relier le label d'un bag à l'existence d'une structure spécifique de ses instances (*concept*). L'article fondateur du MIL [Dietterich et al., 1997] propose l'hypothèse standard selon laquelle un bag est positif si au moins une de ses instances est positive. Cette hypothèse a été élargie et assouplie pour s'adapter à des problèmes nécessitant des relations plus complexes entre instances et bags [Foulds and Frank, 2010].

Les hypothèses de MIL utilisées dans la littérature dépendent du style de problème à résoudre et de l'application, mais nous pouvons distinguer principalement deux catégories de méthodes : 1) Les méthodes locales, qui classent d'abord les instances avant d'agréger leurs résultats pour prédire les labels des bags. 2) Les méthodes globales, qui considèrent directement les bags comme des entités uniques (par exemple, des distributions ou des nuages de points), définissent des métriques adaptées [Cheplygina et al., 2015] et utilisent des méthodes classiques comme les SVM [Andrews et al., 2002] pour les classer.

Le MIL trouve des applications variées, telles que l'analyse moléculaire et la classification

d'images [Dietterich et al., 1997, Andrews et al., 2002] et particulièrement au cours de ces dix dernières années, l'imagerie médicale [Cheplygina et al., 2019]. Ce gain d'intérêt est essentiellement dû à la volonté d'apprendre des concepts visuels complexes à partir de données peu annotées.

En cytométrie classique, [Laziri et al., 2023] proposent des méthodes de réduction de dimension à base de géométrie de l'information à partir de la mesure de quatre variables pour faire du clustering de patients sans visée prédictive. [Hacquard et al., 2023] ont défini un modèle de classification de mesures qu'ils ont appliqué à la classification de patients. D'autres se penchent vers une approche d'imagerie [Sahasrabudhe et al., 2020].

En considérant les patients comme des bags et les cellules comme des instances de ces bags, nous proposons d'utiliser le paradigme du MIL pour la classification de patients en cytométrie. Nous commencerons par illustrer la méthode en modélisant la distribution des cellules d'un bag par un simple modèle de mélange gaussien à deux composantes, avec une proportion variable selon les bags. Afin de capturer la complexité des données, nous introduirons ensuite une méthode qui repose sur l'extraction d'une structure sous-jacente des cellules à l'aide d'un auto-encodeur, structure exploitée par un mélange d'experts. Enfin, nous comparerons les performances des méthodes proposées à celles de méthodes existantes.

2 Modèles

A partir de données de cytométrie d'échantillons sanguins provenant de Metafora, nous souhaitons d'une part, pouvoir prédire l'état de santé d'un patient (positif (1) ou négatif (0) à la maladie), d'autre part déterminer des caractéristiques générales des cellules tumorales. Chaque patient est considéré comme un bag, étiqueté par son état de santé. Nous disposons de plus, pour quelques patients malades, de l'étiquetage (sain, tumoral) des cellules, instances du bag. Nous proposons deux modèles: l'un basé sur un gaussien mélange simple, pour fixer les ingrédients d'une méthode de MIL locale, l'autre entraînant un auto-encodeur pour capturer des patterns sous-jacents.

2.1 EM-MIL: un modèle de MIL par mélange simple

Nous considérons que chaque patient est un bag dont les cellules sont issues d'un modèle de mélange de deux gaussiennes, l'une représentant la composante des cellules saines, l'autre celle des cellules tumorales. La définition des composantes est partagée d'un bag à l'autre, la proportion dépendant du bag. Les labels des instances des bags sans annotations locales sont alors traités comme des variables latentes. Les probabilités a posteriori des instances sont ensuite utilisées pour la construction de la règle de décision au niveau du bag. C'est donc une méthode locale qui définit la classification d'un bag à partir de ses instances. Cette méthode très basique permet de dérouler l'ensemble des ingrédients du MIL. Elle étend les travaux de [Li et al., 2013] qui proposent une méthode de MIL par modèle de mélange de

contamination, mais avec la même proportion dans les différents bags.

De façon plus formelle, soit un ensemble de n bags étiquetés $\{(X_i, Y_i), 1 \leq i \leq n\}$. L'étiquette $Y_i \in \{0, 1\}$ est binaire et chaque bag $X_i \in \mathbb{R}^{m_i \times d}$ possède m_i instances $X_{ij} \in \mathbb{R}^d, j = 1, \dots m_i$. Nous supposons l'indépendance entre les bags d'une part, entre les instances d'un même bag d'autre part et posons un modèle de mélange à deux composantes gaussiennes $(1 - \pi_i)f_0 + \pi_i f_1$ où $f_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$ représente la composante des cellules saines et $f_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ celle des cellules tumorales, π_i étant la proportion de cellules tumorales. Pour un patient sain, $\pi_i = 0$. Lorsque les cellules d'un bag ne sont pas annotées, l'appartenance de la cellule j du bag i à l'une ou l'autre des composantes est gérée par la variable latente $Z_{ij} \in \{0, 1\}$ telle que $P(Z_{ij} = 1) = \pi_i$. Dans le cas contraire, cette variable est connue et sa valeur est directement utilisée (apprentissage semi-supervisé). Les paramètres sont estimés par un algorithme EM [Dempster et al., 1977]. L'étape E consiste à calculer l'espérance des variables latentes conditionnellement aux observations et pour une valeur courante du paramètre. De façon classique:

$$t_{ij}^{(c)} = \mathrm{P}(Z_{ij} = 1 \mid X_{ij}; \Theta^{(c)}) = \frac{\pi_i^{(c)} f_1(X_{ij}; \theta_1^{(c)})}{\pi_i^{(c)} f_1(X_{ij}; \theta_1^{(c)}) + (1 - \pi_i^{(c)}) f_0(X_{ij}; \theta_0^{(c)})}.$$

À l'étape M, les proportions sont estimées par $\pi_i^{(c)} = \sum_{j=1}^{m_i} t_{ij}^{(c)} / m_i$ pour les bags de cellules non annotées et par la proportion de cellules tumorales pour les bags de cellules annotées. Les autres paramètres sont mis à jour avec les formules classiques, où les $t_{ij}^{(c)}$ sont remplacés par les valeurs observées des labels dans le cas de cellules annotées. Les t_{ij} calculés sur toutes les cellules, annotées ou pas, sont ensuite utilisés pour définir une règle de décision de prédiction du label d'un bag. Il en existe plusieurs sortes [Li et al., 2013].

La règle du produit. Elle découle directement de la règle de Bayes et d'une factorisation de la probabilité jointe. Notons α la proportion des bags positifs dans le jeu d'entraînement.

$$\hat{Y}_i = \begin{cases} 1 & \text{si } \alpha^{(1-m_i)} \prod_j [\mathrm{P}(Z_{ij} = 1|X_{ij}) + \frac{\alpha - \pi_i \alpha}{1 - \pi_i \alpha} \mathrm{P}(Z_{ij} = 0|X_{ij})] \geq \\ & (1 - \alpha)^{(1-m_i)} \prod_j \frac{1-\alpha}{1-\pi_i \alpha} \mathrm{P}(Z_{ij} = 0|X_{ij}) \\ 0 & \text{sinon} \end{cases}$$

Bien que théoriquement justifiée, la règle du produit est très sensible à l'accumulation des erreurs d'estimation de probabilité. À la place, la règle de la somme, plus robuste aux erreurs, lui est préférée.

La règle de la somme. [Li et al., 2013] montre que la règle de la somme est obtenue par une approximation de la règle du produit. Elle revient à utiliser une moyenne pondérée des probabilités.

$$\hat{Y}_i = \begin{cases} 1 & \text{si } \alpha(1 - m_i) + \sum_j \mathrm{P}(Z_{ij} = 1|X_{ij}) + \frac{\alpha - \pi_i \alpha}{1 - \pi_i \alpha} \mathrm{P}(Z_{ij} = 0|X_{ij}) \geq \\ & (1 - \alpha)(1 - m_i) + \sum_j \frac{1-\alpha}{1-\pi_i \alpha} \mathrm{P}(Z_{ij} = 0|X_{ij}) \\ 0 & \text{sinon} \end{cases}$$

La règle gamma. Lorsque la règle de décision des instances détermine directement le label prédit $\hat{Z}_{ij}$ par 0 ou 1 sans passer par la modélisation du score $\mathrm{P}(Z_{ij}|X_{ij})$, [Li et al., 2013] proposent de remplacer le score par le label prédit dans la règle somme. En posant $\gamma_i = \sum_j \mathbf{1}_{\{\hat{Z}_{ij}=1\}}/m_i$, la proportion d'instances prédites positives dans un bag X_i, la règle gamma s'écrit alors:

$$\hat{Y}_i = 1 \text{ si } \gamma_i \geq \frac{(1-\pi\alpha)(1-2\alpha)}{2(1-\alpha)m_i} + \pi\alpha; \quad \hat{Y}_i = 0 \text{ sinon.}$$

2.2 Auto-encodeurs et Mélange d'experts

Le modèle de mélange simple à deux composantes reste limité par l'hypothèse (trop simple) de distributions gaussiennes. Pour dépasser cette limitation, nous explorons une approche plus flexible en utilisant des auto-encodeurs pour extraire une représentation pertinente des cellules, capturant ainsi des caractéristiques plus complexes. Nous appliquons ensuite un mélange d'experts sur cette structure latente pour améliorer la classification des patients.

Auto-encodeurs

Un auto-encodeur est un réseau de neurones conçu pour encoder les données d'entrée dans une représentation compressée, puis les décoder pour reconstruire une représentation de sortie aussi similaire que possible à l'entrée. Le réseau se compose principalement de l'encodeur, qui compresse l'entrée, et du décodeur, qui la reconstruit. Les auto-encodeurs sont entraînés de manière non supervisée, ce qui leur permet d'apprendre des représentations informatives des données sans avoir besoin d'étiquettes.

Il existe plusieurs types d'auto-encodeurs dans la littérature. Pour nos expériences, nous avons utilisé l'auto-encodeur orthogonal (OAE) [Wang et al., 2019] qui permet de projeter les données sur des plans orthogonaux. Le problème d'optimisation s'écrit :

$$\min_{\theta,\theta'} \|X - \hat{X}\|_F + \lambda \|H^T H - I\|_F^2$$

où $f_\theta(X) = H$ est le réseau de neurones représentant la fonction d'encodage et $g_{\theta'}(H) = \hat{X}$ celui représentant la fonction de décodage. Le paramètre λ permet l'expression de la contrainte d'orthogonalité.

Mélange d'experts

Le mélange d'experts (Mixture of Experts, MoE, [Gormley and Frühwirth-Schnatter, 2019]) est une approche d'apprentissage où plusieurs modèles spécialisés (experts) collaborent pour résoudre un problème complexe. Plutôt que d'utiliser un modèle unique, le MoE divise l'espace des données en sous-régions, chaque expert Φ_k se spécialisant sur une partie spécifique du problème. Dans notre cas, nous appliquons le MoE aux groupes de cellules révélés par la représentation latente de l'auto-encodeur. Nous attribuons à chaque expert un poids w_k qui correspond à la proportion du groupe de cellules correspondant, reflétant ainsi son importance relative dans la classification. Cette approche permet une modélisation plus flexible en tenant compte de la complexité des sous-populations cellulaires et de leur impact sur l'état de gravité du patient. Nous proposons donc un processus en quatre étapes principales :

1. Extraction d'une représentation latente H : un auto-encodeur projette les cellules dans un espace latent réduit à trois dimensions. Cette représentation permet de capturer les structures sous-jacentes complexes des données.

2. Clustering des cellules : dans cet espace latent, un modèle de mélange gaussien à K composantes regroupe les cellules en sous-populations homogènes avec des proportions ou poids w_k respectifs.

3. Classification par mélange d'experts : à chaque groupe de cellules est associé à un expert Φ_k, modélisé ici par un ensemble d'arbres de décision (Gradient Boosting). Pour entraîner ces experts, le label Z_{ij} attribué à chaque cellule X_{ij} est le label observé pour les cellules provenant de bags de cellules annotées, l'état de gravité Y_i du patient si ce n'est pas le cas. Le poids de chaque expert dans le mélange est déterminé par la proportion de cellules appartenant à son groupe. Le classifieur final pour une cellule est donc le suivant :
$$\Phi^{(\mathrm{MoE})}(X_{ij}) = \sum_{k=1}^{K} w_k \, \Phi_{\mathrm{k}}(H_{ij})$$
où $H_{ij} = f_\theta(X_{ij})$ est la représentation de l'instance X_{ij} dans l'espace de l'auto-encodeur.

4. Prédiction de l'état d'un nouveau patient : les cellules sont d'abord projetées dans l'espace latent par l'auto-encodeur. Ensuite, chaque expert prédit l'état de gravité de toutes les cellules et la prédiction finale pour une cellule est calculée comme la moyenne pondérée des prédictions des experts. Enfin, l'état global du patient est obtenu en utilisant les règles de décision définies en section 2.1.

3 Expérimentations numériques

Dans cette communication, nous présenterons les résultats de comparaison de ces méthodes sur des données simulées comme sur des données réelles. En particulier, nous montrerons la robustesse de la règle somme, même si elle n'est pas la règle théoriquement optimale. Sur un jeu de données d'une étude sur le Myélome multiple issu du laboratoire de Metafora, nous montrerons l'apport de la représentation par auto-encodeur, faisant apparaître une structure intéressante. Enfin, nous challengerons les méthodes locales présentées avec une méthode globale alliant MIL, réseaux de neurones et mécanisme d'attention, proposé par [Ilse et al., 2018].

References

[Andrews et al., 2002] Andrews, S., Tsochantaridis, I., and Hofmann, T. (2002). Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 15.

[Carbonneau et al., 2018] Carbonneau, M.-A., Cheplygina, V., Granger, E., and Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.

[Cheplygina et al., 2019] Cheplygina, V., De Bruijne, M., and Pluim, J. P. (2019). Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296.

[Cheplygina et al., 2015] Cheplygina, V., Tax, D. M., and Loog, M. (2015). Multiple instance learning with bag dissimilarities. *Pattern recognition*, 48(1):264–275.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

[Dietterich et al., 1997] Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71.

[Foulds and Frank, 2010] Foulds, J. and Frank, E. (2010). A review of multi-instance learning assumptions. *The knowledge engineering review*, 25(1):1–25.

[Gormley and Frühwirth-Schnatter, 2019] Gormley, I. C. and Frühwirth-Schnatter, S. (2019). Mixture of experts models. In *Handbook of mixture analysis*, pages 271–307. Chapman and Hall/CRC.

[Hacquard et al., 2023] Hacquard, O., Blanchard, G., and Levrard, C. (2023). Statistical learning on measures: an application to persistence diagrams. *arXiv preprint arXiv:2303.08456*.

[Ilse et al., 2018] Ilse, M., Tomczak, J., and Welling, M. (2018). Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR.

[Laziri et al., 2023] Laziri, W., Allemand, F., Gégout-Petit, A., and Wantz-Mézières, S. (2023). Méthodes de réduction de dimension basées sur l'algorithme fine pour le clustering de patients à partir de données de cytométrie en flux. In *54es Journées de Statistiques de la SFDS*.

[Li et al., 2013] Li, Y., Tax, D. M., Duin, R. P., and Loog, M. (2013). Multiple-instance learning as a classifier combining problem. *Pattern recognition*, 46(3):865–874.

[Sahasrabudhe et al., 2020] Sahasrabudhe, M., Sujobert, P., Zacharaki, E. I., Maurin, E., Grange, B., Jallades, L., Paragios, N., and Vakalopoulou, M. (2020). Deep multi-instance learning using multi-modal data for diagnosis of lymphocytosis. *IEEE Journal of Biomedical and Health Informatics*, 25(6):2125–2136.

[Wang et al., 2019] Wang, W., Yang, D., Chen, F., Pang, Y., Huang, S., and Ge, Y. (2019). Clustering with orthogonal autoencoder. *IEEE Access*, 7:62421–62432.

Leveraging shared information in multimodal generative modeling

Agathe Senellart [1] & Stéphanie Allassonnière [2]

[1] *(UMR1346, Université Paris-Cité, Inria, Inserm), France, agathe.senellart@inria.fr*
[2] *(UMR1346, Université Paris-Cité, Inria, Inserm), France, stephanie.allassonniere@u-paris.fr*

Résumé.

Dans ce travail, nous nous intéressons à la modélisation jointe de plusieurs modalités à l'aide d'autoencodeurs variationnels. Plus précisément, pour X aléatoire contenant M modalités, $X = (x_1, x_2 \dots x_M)$, nous souhaitons modéliser la distribution jointe $p_\theta(X)$ et les distributions conditionnelles $p_\theta(x_i|(x_j)_{j\in S})$ permettant d'inférer une modalité manquante x_i, $1 \leq i \leq M$ à partir d'un sous-ensemble S de modalités disponibles. Ces dernières années, plusieurs adaptations du VAE ont été proposées pour le cadre multimodal. Dans de nombreux modèles, les modalités sont aggrégées avec une opération simple comme un Produit d'Experts, une mixture d'Experts ou une Mixture de Produit d'experts. Néanmoins des travaux récents suggèrent que cette approche présente des limitations théoriques qui limitent les capacités de modélisation des modèles dans des cas d'interactions complexes. Dans ce travail, nous proposons donc une approche alternative pour modéliser les distributions jointe et conditionnelles entre les modalités. La construction de notre modèle se découpe en trois étapes: 1) Dans un premier temps, nous modélisons la distribution jointe des données à l'aide d'un VAE simple. 2) Dans un second temps, nous proposons d'extraire l'information sémantique commune aux différentes modalités à l'aide d'une méthode non supervisée. Cette étape permet d'isoler l'information partagée entre les modalités qui sera pertinente pour modéliser les distributions conditionnelles. 3) Enfin nous modélisons les distributions conditionnelles $p(x_i|x_j)$ à l'aide de flots génératifs et en s'appuyant sur la modélisation jointe réalisée à l'étape 1) et l'information partagée entre les modalités extraite à l'étape 2). Notre méthode obtient des résultats compétitifs sur plusieurs jeux de données.

Mots-clés. Variational Autoencoder, Self-Supervised Learning, Multimodality, ...

Abstract. From medical diagnosis to autonomous vehicles, critical applications rely on the integration of multiple heterogeneous data modalities. Multimodal Variational Autoencoders offer versatile and scalable methods for generating unobserved modalities from observed ones. More precisely, if X is a multimodal random observation with M modalities, $X = (x_1, x_2 \dots x_M)$, we aim to model the joint distribution $p_\theta(X)$ and conditional distributions $p_\theta(x_i|(x_j)_{j\in S})$ to generate new samples. Recent models using mixtures-of-experts or product-of-experts aggregation suffer from theoretically grounded limitations that caps their generation quality on complex datasets. In this article, we propose a novel interpretable model able to learn all the joint and conditional distributions without introducing mixture aggregation. Our method follows a multistage training process: 1) First modelling the joint

distribution with variational inference 2) Extracting the information shared accross modalities using a self-supervised method 3) Modelling the conditional distributions $p(x_i|x_j)$ with Normalizing Flows while leveraging the shared information from step 2) and the joint model learned in step 1). Our method achieves competitive results on several benchmark datasets.

Keywords. Keyword 1, keyword 2, ...

1 Méthode

Nous supposons que nous observons des échantillons multimodaux $X = (x_1, x_2, \ldots, x_M)$ avec M modalités à partir d'une distribution inconnue $p(X)$. Nous cherchons à approximer cette distribution ainsi que les distributions conditionnelles avec des distributions paramétriques $p_\theta(X)$, $p_\theta(x_j|x_i)$ pour tout $1 \leq i \neq j \leq M$.

Dans le cadre du VAE, nous supposons qu'il existe une représentation latente partagée z, à partir de laquelle toutes les modalités peuvent être générées avec des distributions paramétriques $(p_\theta(x_j|z))_{1\leq j\leq M}$ appelées *decodeurs*. Par exemple, pour une modalité d'image x_1, $p_\theta(x_1|z)$ peut être une distribution gaussienne $\mathcal{N}(\mu_\theta(z), \Sigma_\theta(z))$ dont la moyenne et la variance sont données par un réseau de neurones. Dans la plupart des cas [10, 9], chaque modalité est supposée être conditionnellement indépendante des autres étant donné z, de sorte que le modèle joint s'écrit :

$$p_\theta(X, z) = p_\theta(X|z)p(z) = p(z)\prod_{j=1}^{M} p_\theta(x_j|z)\,, \tag{1}$$

$p(z) = \mathcal{N}(z; 0, I)$ est une distribution *a priori*. θ correspond aux paramètres de tous les décodeurs.

Notre méthode se découpe en trois étapes distinctes :

- Dans un premier temps, nous apprenons le paramètre θ avec de l'inférence variationnelle afin de modéliser la distribution jointe.
- Dans un second temps, nous cherchons à extraire l'information partagée entre les modalités qui sera pertinente pour approximer les distributions conditionnelles $p(x_i|x_j)$.
- Enfin nous modélisons les distributions conditionnelles $p(x_i|x_j)$ avec des flots génératifs et en nous appuyant sur les deux premières étapes.

1.1 Estimer le paramètre génératif θ

Étant donné N observations multimodales $(X^{(i)})_{1\leq i\leq N}$, un objectif naturel pour estimer θ est d'optimiser la log-vraisemblance des données [3] :

$$\theta^* \in \operatorname*{argmax}_\theta \sum_{i=1}^{N} \log p_\theta(X^{(i)}) = \operatorname*{argmax}_\theta \sum_{i=1}^{N} \left(\log \int_z p_\theta(X^{(i)}, z)dz\right)\,.$$

Comme cet objectif est intractable, on peut recourir à l'inférence variationnelle [3] en introduisant une distribution paramétrique auxiliaire $q_\phi(z|X)$ nous permettant d'obtenir une borne inférieure pour $\log p_\theta(X)$, appelée en anglais "Evidence lower bound" (ELBO).

$$\log p_\theta(X) \geq \mathbb{E}_{q_\phi(z|X)}\left[\log p_\theta(X|z)\right] - KL(q_\phi(z|X)||p_\theta(z)) = \mathcal{L}(X;\theta,\phi)\,.$$

Cette borne est analytiquement calculable et peut être optimisée par descente de gradient stochastique [3]. La distribution $q_\phi(z|X)$ est généralement appelée *encodeur* et on peut montrer qu'elle approxime la vraie postérieure $p_\theta(z|X)$. [3]

Dans un certain nombre de travaux récents, $q_\phi(z|X)$ est modélisée comme une aggrégation d'encodeurs unimodaux par exemple avec un Produit d'Experts: $q_\phi(z|X)\prod_i q_\phi(z|x_i)$ ou Mixture d'Experts. Nous choisissons de modéliser $q_\phi(z|X)$ comme une distribution gaussienne $\mathcal{N}(\mu_\phi(X), \Sigma_\phi(X))$ dont les paramètres sont donnés par un réseau de neurones prenant toutes les modalités X en entrée. Ce choix permet une modélisation moins restrictive des interactions entre les modalités.

À l'issue de cette étape, nous avons donc estimé le paramètre θ des décodeurs $p_\theta(x_j|z)$. Pour pouvoir générer une modalité x_i à partir d'une autre x_j, $1 \leq i \neq j \leq M$, il nous faut encore modéliser la distributions postérieures $p_\theta(z|x_j)$ qui n'est pas calculable analytiquement. On pourra alors générer x_i en échantillonnant $z \sim p_\theta(z|x_j)$ puis $x_i \sim p_\theta(x_i|z)$.

1.2 Extraire l'information partagée entre les modalités

Dans de nombreux jeux de données multimodales, il existe une certaine quantité d'information sémantique partagée entre les modalités. Par exemple, dans le jeux de données MNIST-SVHN [7], le contenu sémantique partagé est le chiffre présent dans les deux images. L'information d'arrière-plan est *spécifique à la modalité* dans le sens où elle *n'affecte pas* l'autre modalité.

Formellement, supposons que pour tout $1 \leq j \leq M$ nous ayons un projecteur g_j vers un espace de plus petite dimension tel que :

$$\forall 1 \leq i \leq M, p_\theta(x_i|x_j) = p_\theta(x_i|g_j(x_j))\,. \tag{2}$$

Dans ce cas, pour générer la modalité x_i à partir de la modalité x_j, nous pouvons apprendre à approximer $p_\theta(z|g_j(x_j))$ qui est éventuellement plus simple à approximer que $p_\theta(z|x_j)$ si nous utilisons de bons projecteurs.

Comment extraire l'information partagée entre les modalités ? Comment pouvons-nous apprendre des fonctions pertinentes $(g_j)_{1\leq j\leq M}$ préservant l'information partagée entre les modalités tout en éliminant l'information inutile ? Dans nos expériences, nous avons essayé deux méthodes générales : Deep Canonical Correlation Analysis (DCCA) [1] et Contrastive Learning (CL) [5, 6]. Nous *conjecturons* qu'en utilisant ces méthodes, nous pouvons extraire des statistiques pertinentes $(g_j(x_j))_{1\leq j\leq M}$ vérifiant l'équation (2) et nous vérifions cette hypothèse dans nos expériences.

1.3 Modéliser les distributions postérieures

Pour approcher les distributions postérieures $p_\theta(z|g_j(x_j))$ qui sont intractables, nous introduisons des distributions auxiliaires $q_{\phi_j}(z|g_j(x_j))$ et nous optimisons l'objectif suivant utilisé dans [9]:

$$\mathcal{L}_{uni}(X;(\phi_j)_{1\leq j\leq M}) = \sum_{j=0}^{M} KL(q_\phi(z|X)|q_{\phi_j}(z|g_j(x_j)))\,. \tag{3}$$

Intuitivement, minimiser (3) encourage $q_{\phi_j}(z|g_j(x_j))$ à recouvrir tous les modes ou supports de $q_\phi(z|X)$ (qui, on le rappelle, est une approximation de $p_\theta(z|X)$) pour différentes valeurs de X associées à une même information partagée $g_j(x_j)$.

Pour une modélisation expressive, nous choisissons de modéliser $q_{\phi_j}(z|g_j(x_j))$ avec des flots génératifs(Normalizing Flows (NF) en anglais). Cet outil de modélisation puissant, permet de modéliser des distributions complexes et différentiables [4]. Un flot est une transformation inversible f qui peut être appliquée à une distribution initiale pour en créer une nouvelle plus complexe.

Pour $1 \leq j \leq M$, nous utilisons comme distribution initiale pour $q_{\phi_j}(z|g_j(x_j))$ une loi normale et appliquons ensuite 2 flots MAF (Masked Autoregressive flows) [4].

À l'issue de cette étape, nous avons donc modélisé les postérieures $q_{\phi_j}(z|g_j(x_j))$ nous permettant de faire de l'inférence conditionnelle d'une modalité à une autre. À cause de l'entraînement Joint des flots génératifs (NF) à cette étape, nous appelons notre méthode JNF-Shared. Nous appelons JNF, la variante de notre méthode où l'étape 2) n'est pas réalisée. Autrement dit on ne cherche pas à extraire l'information partagée et prenons simplement $g_j(x_j) = x_j$.

1.4 Distributions conditionnelles à un sous-ensemble de modalités

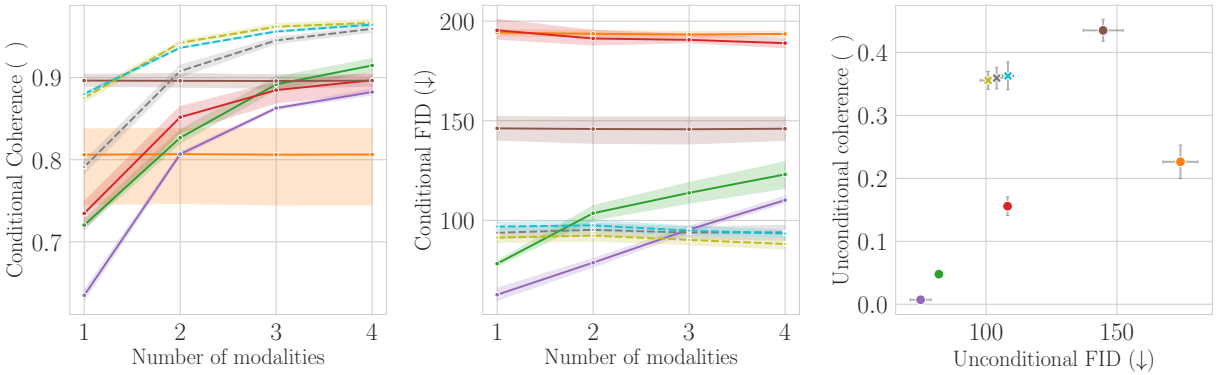
Si nous disposons d'un sous-ensembles de modalités $S \subset [|1, M|]$ et que nous souhaitons inférer une modalité manquante, nous avons besoin d'approximer $p(z|(x_j)_{j\in S})$. Pour ce faire, nous utilisons un produit d'experts:

$$p(z|(x_j)_{j\in S}) \approx \frac{1}{Z}\frac{\prod_{j\in S} q_{\phi_j}(z|g_j(x_j))}{p(z)^{|S|-1}}\,, \tag{4}$$

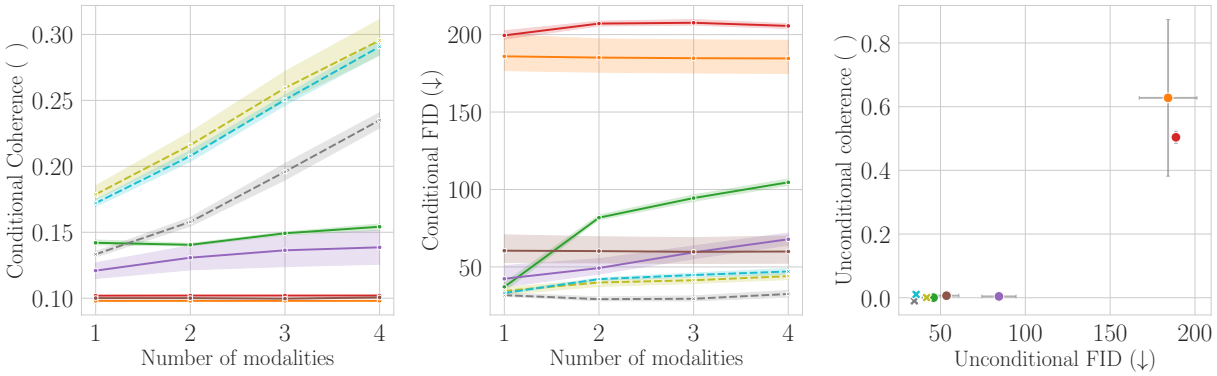
où Z est une constante de normalisation. Ce produit d'experts émerge naturellement de l'équation (1) [10]. Nous utilisons ce produit d'experts (PoE) seulement pendant l'inférence et pas pendant l'entraînement ce qui permet d'éviter les limitations observées liées à l'utilisation d'un PoE ([8]).

2 Expériences

Nous testons notre méthode sur kes jeux de données PolyMNIST et Translated PolyMNIST [2]. Nous évaluons la *cohérence* des générations (est-ce que les données générées dans différentes modalités partagent le même contenu sémantique ?) et la diversité des générations en calculant un score de FID (Fréchet Inception Distance) entre les données réelles et les données générées.



(a) PolyMNIST



(b) Translated PolyMNIST

Figure 1: Pour chaque jeu de données : dans les deux figures de gauche, nous présentons des résultat de générations conditionelles en variant le nombre de modalités en entrée. Dans la figure de droite, nous présentons des résultats de génération jointe. Nos modèles obtiennent les meilleures cohérences tout en maintenant de faibles FID.

Bibliographie

References

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep Canonical Correlation Analysis. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1247–1255. PMLR, May 2013. ISSN: 1938-7228.

[2] Imant Daunhawer, Thomas M. Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E Vogt. On the limitations of multimodal VAEs. In *International Conference on Learning Representations*, 2022.

[3] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, May 2014. arXiv:1312.6114 [cs, stat].

[4] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked Autoregressive Flow for Density Estimation. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[5] Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S. Melo, Ana Paiva, and Danica Kragic. Geometric Multimodal Contrastive Representation Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 17782–17800. PMLR, June 2022. ISSN: 2640-3498.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[7] Yuge Shi, N. Siddharth, Brooks Paige, and Philip H. S. Torr. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. *arXiv:1911.03393 [cs, stat]*, November 2019. arXiv: 1911.03393.

[8] Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Generalized Multimodal ELBO. *ICLR*, 2021.

[9] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint Multimodal Learning with Deep Generative Models. *arXiv:1611.01891 [cs, stat]*, November 2016. arXiv: 1611.01891.

[10] Mike Wu and Noah Goodman. Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Références

Apprentissage par méthode bayésienne

HETEROGENEOUS BAYESIAN TRANSFER LEARNING FOR COMPLEX ORDINARY DIFFERENTIAL EQUATIONS

Youba Abed[1,2] & Victor Costa[1,4] & Benoit Celse[1,5] & Denis Guillaume[1,6] & Julien Jacques[3]

[1] *IFPEN, Rond-point de l'échangeur de Solaize, France*
[2] *Université Lumière Lyon 2, Université Claude Bernard Lyon 1, ERIC, 69007, Lyon, France, youba.abed@ifpen.fr*
[3] *Université Lumière Lyon 2, Université Claude Bernard Lyon 1, ERIC, 69007, Lyon, France, julien.jacques@univ-lyon2.fr*
[4] *victor.costa@ifpen.fr*
[5] *benoit.celse@ifpen.fr*
[6] *denis.guillaume@ifpen.fr*

Résumé. Le groupe IFP est un leader dans la recherche et la formation en énergie et environnement, notamment dans le développement et la commercialisation de catalyseurs. La création de modèles prédictifs fiables pour ces catalyseurs nécessite des données expérimentales coûteuses. Une première approche de transfert bayésien a permis de réduire les coûts et le temps de modélisation en réutilisant les données existantes, mais elle supposait des structures de modèles identiques. Cet article explore une méthode de transfert bayésien hétérogène pour des modèles de structures différentes, montrant des performances supérieures aux méthodes classiques sans transfert sur des données simulées.

Mots-clés. Transfert Learning Hétérogène, Transfert Learning Hétérogène Bayésien

Abstract. The IFP group is a leader in research and training in the energy and environmental sector, particularly in the development and commercialization of catalysts. Building reliable predictive models for these catalysts requires costly experimental data. An initial Bayesian transfer approach helped to reduce modeling costs and development time by reusing existing data but assumed identical model structures. This article explores a heterogeneous Bayesian transfer learning method for models with different structures, demonstrating better performance than traditional non-transfer methods on simulated data.

Keywords. Heterogeneous Transfer Learning, Heterogeneous Bayesian Transfer Learning

1 The problem

1.1 Context

The IFP Group designs and commercializes catalysts which are essential for refining fuel from oil, especially in hydrotreating and hydrocracking processes. Hydrocracking transforms heavy oil (like vacuum gas oil) or renewable feedstocks (vegetable oils, tyre/plastic oils) into lighter, more valuable fuels like diesel, using heterogeneous catalysts under high

temperature and hydrogen pressure. Since impurities like nitrogen and sulfur are inhibitors for the catalyst, a prior hydrotreatment is required to remove them. Predicting the nitrogen concentration after hydrotreatment is crucial for improving efficiency, and this prediction is usually done using physical models based on material balance described by Ordinary Differential Equations (ODEs). Given a data set represented by the following features: {Temperature ($Temp$), Hydrogen partial pressure (ppH_2), Inverse of the residence time ($LHSV$), Weighted average of the simulated distillation (TMP), Nitrogen content in feedstock (N_0), Sulfur content in feedstock (S_0), Resins content in feedstock(Res_0)}. Kinetic experts provided the following relation between these features and the output nitrogen concentration (y), which is described by the following ODE (Per Julian Becker and Benoit Celse (2024)):

$$\begin{aligned}\frac{dy}{dt} = &- k_0 \frac{exp\left(-\frac{E_a}{R_g}\left(\frac{1}{T}-\frac{1}{T_{ref}}\right)\right)\left(\frac{ppH_2}{ppH_{2,ref}}\right)^m y^n}{(1+A_0 Res_0)\left(1+\frac{C_0 N_0}{1+S_0}\right)}\times \\ &\left(1-u\cdot exp\left(-\frac{b}{R_g}\left(\frac{1}{T}-\frac{1}{T_{ref}}\right)\right)\left(\frac{ppH_2}{ppH_{2,ref}}\right)^\alpha \left(\frac{TMP}{TMP_{ref}}\right)^v y^l\right)\end{aligned} \tag{1}$$

where $\theta = (k_0, E_a, m_n, \alpha, b, A_0, C_0, u, t, v)$ are the parameters to optimize using the data set, and $R_g, T_{ref}, ppH_{2,ref}, TMP_{ref}$ are fixed references values. As previously mentioned, $LHSV$, the inverse of the residence time, is used in the integration of the ODE as shown below:

$$y = \int_0^{\frac{1}{LHSV}} \left(\frac{dy}{dt}\right) dt \tag{2}$$

This ODE is one of many examples of modeling nitrogen output specifically for fossil feedstocks. The fit of these parameters ODEs requires an important amount of costly experimental data (around 100 points). Every 3 years, a new catalyst is developed. Each catalyst has it's own model with different parameters that need to be refitted. Even if the same ODE (Equation 1) is valid for different catalyst, their parameters θ should be refitted for each new catalyst.

1.2 Problem setup

For non-fossil feedstocks, model predictions using Equation (1) are not accurate. Experimental results suggest that a new feature (Oxygen ($Oxyg$)) is now relevant to the behavior of the output nitrogen content. For this reason, the model was modified to include this new feature and it's corresponding parameter. The new equation is presented in the following ODE:

$$\begin{aligned}\frac{dy}{dt} = &- k_0 \frac{exp\left(-\frac{E_a}{R_g}\left(\frac{1}{T}-\frac{1}{T_{ref}}\right)\right)\left(\frac{ppH_2}{ppH_{2,ref}}\right)^m y^n}{(1+A_0 Res_0)\left(1+\frac{C_0 N_0}{1+S_0}\right)(1+pOxyg)}\times \\ &\left(1-u\cdot exp\left(-\frac{b}{R_g}\left(\frac{1}{T}-\frac{1}{T_{ref}}\right)\right)\left(\frac{ppH_2}{ppH_{2,ref}}\right)^\alpha \left(\frac{TMP}{TMP_{ref}}\right)^v y^l\right)\end{aligned} \tag{3}$$

A new input feature ($Oxyg$) and a new parameter (p) are introduced in the denominator when compared to Equation (1). Note that the common parameters in equations (1) and (3) represent the same effects but have different values. Additionally, the distribution of the common features differs between the two equations. The goal is to estimate the parameters of Equation (3) with minimal cost data. This motivates the idea of leveraging the knowledge from the data set of equation (1) and transferring it to improve this optimization problem. This constitutes a transfer learning approach (Bao et al. (2023)), where knowledge from a source domain (data set and parameters of Equation (1)) is reused to enhance optimization in a target domain (parameters of equation (3)). Our case falls within the category of Heterogeneous Transfer Learning (HTL), as the feature spaces of the two data sets differ. The symbol $\theta_{\mathrm{S}} \in \mathbb{R}^{10}$ represents the source parameters (from Equation 1), while $\theta_{\mathrm{T}} \in \mathbb{R}^{11}$ refers to the target parameters (from Equation 3).

2 Methodologies for fitting parameters

This section presents two methods to fit parameters: Bayesian inference method and the proposed heterogeneous bayesian transfer learning method (HBTL).

2.1 Metropolis-Hastings algorithm

- **Bayesian Inference with Weakly Informative Prior**

  The parameters are fitted similarly to Iapteff et al. (2023) using Bayesian inference by assuming the normality of the residuals:

$$\begin{aligned} y_i &= f_\theta(x_i) + \epsilon_i, \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_i^2), \end{aligned} \tag{4}$$

  where $f_\theta(\cdot)$ is the solution of a differential equation (e.g., Equation 1 or 3) calculated with the Runge-Kutta method. A heteroscedastic model is considered, with the variance of ϵ_i depending on the observations. The expression for σ_i is chosen such that maximizing the likelihood in Equation (4) is equivalent to minimizing the cost function $\sum_{i=1}^{N} \frac{(\hat{y}_i - y_i)^2}{y_i}$ where N is the sample size.

- **Posterior Distribution Calculation using MCMC/MH Algorithm**

  Since the posterior distribution cannot be calculated analytically, the Metropolis-Hastings (MH) within Gibbs algorithm (Tierney, 1994), a type of MCMC algorithm, is employed with a weakly informative prior. The prior is a Gaussian distribution with a large variance: $\mathcal{N}(0_{\mathbb{R}^n}, m \times \mathbb{I}_n)$, where n is the number of parameters and m is a large value, typically 10^{100}. For each parameter, a normal distribution is used as the proposal distribution for sampling. In each iteration, the prediction $\hat{y}_i$ is computed using the Runge-Kutta method to solve the ODE. The proposal distribution is centered on the previous draw, with a standard deviation that can be fixed.

However, as is well known in the MH algorithm, the acceptance ratio should be neither too small nor too high (Gelman et al. (1996)). To achieve this, in our experiments, the standard deviations of the proposal distributions are first initialized to 10% of the initial source parameters values (source parameters are initialized for MCMC) and then adjusted in each iteration according to the following rule: if the acceptance rate is higher than 70%, the standard deviation is increased, as larger standard deviations lead to a lower acceptance ratio; if the acceptance rate is lower than 30%, the standard deviation is decreased, as smaller standard deviations lead to a higher acceptance ratio; and if the acceptance rate is between 30% and 70%, the standard deviation remains unchanged. After the burn-in period, the estimates $\hat{\theta}$ and $\text{Var}(\hat{\theta})$ are obtained from the mean and covariance of the remaining samples, respectively.

2.2 Heterogeneous Bayesian Transfer Learning Method

After estimating $\hat{\theta}_S$ and $\text{Var}(\hat{\theta}_S)$ using the MH method described in Section 2.1, our proposed HBTL method for optimizing the target parameters follows a similar approach but incorporates the following Gaussian prior:

$$\tilde{\pi}(\theta_T) \sim \mathcal{N}\left(\begin{pmatrix}\hat{\theta}_{\text{S}} \\ 0\end{pmatrix}, g \times \begin{pmatrix}\text{Var}(\hat{\theta}_{\text{S}}) & 0_{\mathbb{R}_{10}} \\ 0_{\mathbb{R}_{10}}^{\top} & m\end{pmatrix}\right), \tag{5}$$

where g is a hyperparameter that controls the importance of this prior, and m is a similar value used for the prior in Section 2.1 (10^{100}), reflecting our lack of information about the parameter p.

3 Methods comparison

The evaluation of the two suggested methods was conducted on simulated data sets from equations (1) and (3). Another baseline was provided.

3.1 Synthetic Data Sets

The source parameters were fixed to generate the synthetic data sets. The target parameters shared with the source (parameters common between Equations 1 and 3) where obtained by increasing the source parameters by 20% of their respective values: $\theta_{i\text{T}} = 1.2 \times \theta_{i,\text{S}}$ for $i = 1, 2, \ldots, 10$. For simplicity, it was assumed that p is equal to A_0, as both A_0 and p appear in the denominator of Equations (1) and (3). Then, the data sets were generated using the Kennard-Stone algorithm (R. W. Kennard and L. A. Stone (1969)). The choice of the Kennard-Stone algorithm ensures that the data sets contain as much information as possible about different samples, preventing the methods from being limited by the data sets' information. Twenty different data sets were simulated for both the source and target domains, with 100 points per data set for the source, 5 to 25 points for the target training data set and around 10,000 points for the target test data set.

Uniform noise $U[0, 0.15\, y_i]$ was added to each output y_i, meaning that for each y_i, a random value uniformly distributed between 0 and 15% of its value was added. This reflects the analytical uncertainty of the output nitrogen measurement.

3.2 Experiments setup

For each target training data set, the two methods, MH and HBTL, are compared on the target test data set using the RMSE (Root Mean Square Error) metric, which measures the difference between the true outputs and the predicted ones. The MH method is trained only on the target training data sets. Different values of g (see Section 2.2) within $\{1, 10, 100, 1000, 10000\}$ and target training sample sizes are used. A baseline is also provided using the well-known gradient descent method (L-BFGS-B algorithm, Byrd et al., 1995), which is trained solely on the target training data sets. For L-BFGS-B, 1000 different initializations are used.

4 Results

To make the comparisons easier to understand, the mean and confidence intervals (using the 2.5th and 97.5th percentiles) were calculated for our HBTL method and the baseline methods (MH and L-BFGS-B). During training, the optimal value of g for our approach was determined to be 10000, as it corresponded to the lowest RMSE observed on the training data sets. Figure 1 shows the evolution of the averaged RMSE on the test target data sets as a function of the target training sample size. It is clear that our HBTL (represented by the black plot) outperforms both approaches without transfer learning, and also exhibits tighter confidence intervals, indicating more stable and reliable predictions. For a target training sample size of 15 points, our approach exhibits an RMSE more than two times smaller than the baseline methods. The gradient descent exhibits very unstable results, which can be explained by the fact that it often gets stuck in local minima, even with 1000 different initializations, as the models used (ODEs) are very complex. In contrast, the MH yields better results, which can be attributed to the tuning of the standard deviations of the proposal distribution, giving it an advantage over gradient descent.

5 Conclusion and perspectives

This study emphasizes the importance of transfer learning and demonstrates that our HBTL approach is effective. However, its main drawback is that it assumes knowledge of the model structure (the target's ODE), which is particularly challenging for some new type feedstocks. Our future work focuses on developing HTL methods that do not rely on knowing the target ODE structure.

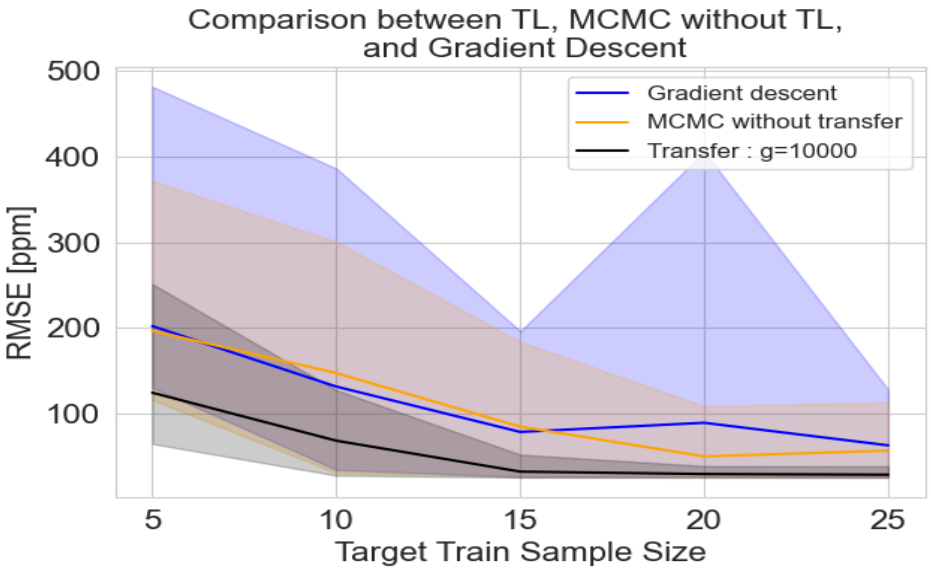


Figure 1: Comparison between HBTL (black plot), MH (orange plot), and L-BFGS-B (blue plot). The bands represent the confident intervals.

References

Per Julian Becker and Benoit Celse (2024), Stepwise Parameter Fitting to Combine Industrial and Pilot Plant Data sets, Computer Aided Chemical Engineering,Volume 53, pp. 901-906.

Runxue Bao, Yiming Sun, Yuhe Gao, Jindong Wang, Qiang Yang, Haifeng Chen, Zhi-Hong Mao, and Ye Ye (2024), A Recent Survey of Heterogeneous Transfer Learning, url: https://arxiv.org/abs/2310.08459.

Loïc Iapteff, Julien Jacques, Benoit Celse and Victor Costa (2023), Reducing the Number of Experimental Points to Fit Kinetic Models: A Bayesian Approach. In: Industrial Engineering Chemistry Research 62.28, pp. 10903–10914.

Luke Tierney (1994), Markov Chains for Exploring Posterior Distributions. In: Biometrika 57.1, pp. 97–109.

Andrew Gelman, Gareth Roberts, and Walter Gilks (1996), Efficient Metropolis Jumping Rules, Institute of Mathematical Statistics, the Annals of Statistics, pp. 1701 - 1728.

Piotr Suder, Jason Xu, and David Dunson (2023), Bayesian Transfer Learning, url: https://arxiv.org/abs/2312.13484.

Richard Kennard and Lloyd Stone (1969), Computer Aided Design of Experiments". In: Technometrics Vol. 11, pp. 137-148.

Richard Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu (1995), A Limited Memory Algorithm for Bound Constrained Optimization, In: Biometrika 57.1, pp. 97–109.

Réseaux de neurones bayésiens pour la métamodélisation des simulateurs numériques

Oumar Baldé [1], Amandine Marrel [1,2] & Georis Billo [1]

[1] *CEA, DES, IRESNE, DER, SESI, Cadarache, 13108 Saint-Paul-Lez-Durance, France;*
E-mail : oumar.balde@cea.fr
[2] *LMA Université d'Avignon, France*

Résumé. La simulation numérique est largement utilisée pour modéliser et prédire le comportement de systèmes physiques complexes. Certains simulateurs peuvent nécessiter un temps CPU important, ce qui limite le nombre de simulations possibles et *de facto* leur utilisation directe et intensive pour réaliser certaines tâches telles que la propagation des incertitudes ou la calibration de paramètres. Une solution est alors d'approcher la sortie du simulateur par un métamodèle, basé sur des techniques d'apprentissage automatique. Les réseaux de neurones bayésiens (RNB) sont des métamodèles théoriquement séduisants car ils combinent la flexibilité des réseaux de neurones artificiels et fournissent une distribution prédictive, quantifiant l'incertitude de prédiction. Cependant, leur mise en œuvre soulève plusieurs questions dont le choix des distributions *a priori* pour les paramètres du réseau. Pour apporter des premiers éléments de réponse, nous étudions ici différentes distributions (issues de la littérature ou hybridées) et illustrons, sur un exemple analytique 1D, leur impact sur la précision de la distribution prédictive du RNB. Les résultats montrent qu'il est possible d'obtenir avec les RNB et certains *a priori* des intervalles de crédibilité plus précis qu'avec d'autres métamodèles de référence (régression par processus gaussien, e.g.). Cela ouvre la voie à des études plus approfondies dans des dimensions supérieures.

Mots-clés. Réseaux de neurones bayésiens, distribution *a priori*, Processus gaussiens, Apprentissage automatique.

Abstract. Numerical simulation is widely used to model and predict the behavior of complex physical systems. Some simulators may require significant CPU time, which limits the number of possible simulations and, consequently, their direct and intensive use for tasks such as uncertainty propagation or parameter calibration. A solution is to approximate the output of the simulator with a metamodel based on machine learning techniques. Bayesian neural networks (BNNs) are theoretically an appealing metamodel because they combine the flexibility of artificial neural networks and provide a predictive distribution that quantifies prediction uncertainty. However, their practical use raises several questions, including the choice of prior distributions for the network parameters. To provide initial answers, we study different distributions (from the literature or hybridized) and illustrate, on a 1D analytical example, their impact on the accuracy of the predictive distribution of BNN. The results show that it is possible to obtain with BNNs and certain priors more accurate credibility intervals than reference methods (Gaussian process regression, e.g.). This paves the way for more in-depth studies in higher dimensions.

Keywords. Bayesian neural networks, *prior* distribiution, Gaussian process, Machine learning.

1 Introduction

Les simulateurs numériques sont devenus des outils incontournables pour modéliser, comprendre et prédire le comportement de systèmes physiques complexes. Cependant, une représentation fidèle de ces phénomènes se fait souvent au prix d'un coût computationnel élevé, limitant ainsi le nombre de simulations possibles. Cela complique et restreint la réalisation de certaines tâches nécessitant de nombreuses simulations, telles que l'étude de la propagation des incertitudes ou la calibration des paramètres incertains (en entrée du simulateur). Pour pallier ce problème, une solution efficace consiste à émuler la sortie du simulateur par un métamodèle basé sur des techniques d'apprentissage automatique [Currin et al., 1991]. Parmi les métamodèles usuels, la régression par processus gaussiens (RPG, [Rasmussen et al., 2006]) se distingue par sa capacité à fournir une distribution de probabilité prédictive, permettant notamment de construire des intervalles de crédibilité associés aux prédictions. Cette caractéristique en fait une référence pour approximer les simulateurs, en particulier dans l'évaluation des risques industriels. Cependant, la RPG présente aussi certaines limites applicatives : ses performances dépendent fortement de certains choix paramétriques, tels que le choix *a priori* d'une fonction de covariance et l'estimation des paramètres associés. En pratique, le choix usuel d'une covariance stationnaire et avec un certain degré de régularité peut s'avérer trop limité pour représenter des phénomènes très complexes, avec des irrégularités locales ou potentiellement chaotiques.

Les réseaux de neurones bayésiens (RNB) offrent alors une alternative intéressante à la RPG, en combinant la flexibilité des réseaux de neurones artificiels et le cadre bayésien qui permet de disposer d'une distribution prédictive. En théorie, les RNB permettent de capturer des relations complexes tout en quantifiant l'incertitude de prédiction. Contrairement à la RPG, ils ne reposent pas sur une hypothèse de stationnarité et peuvent mieux s'adapter aux structures de données complexes. Cependant, leur mise en œuvre soulève plusieurs défis. D'une part, elle nécessite de choisir une architecture neuronale adéquate ainsi que des distributions *a priori* pour les poids et les biais du réseau. D'autre part, l'inférence bayésienne pour les réseaux de neurones est très coûteuse en calcul et requiert souvent des approximations, telles que l'inférence variationnelle ou les méthodes Monte Carlo par Chaînes de Markov (MCMC), qui peuvent introduire des biais et incertitudes supplémentaires. Enfin, bien que les RNB soient théoriquement prometteurs, leurs performances n'ont pas encore été aussi largement et rigoureusement évaluées que celles de la RPG, sur les applications visées.

Ainsi, cette étude explore l'apport des RNB comme alternative à la RPG, en considérant notamment plusieurs choix de distributions *a priori* pour les paramètres du RNB, distributions issues de la littérature ou de l'hybridation de plusieurs approches. Leurs performances sont illustrées sur un exemple analytique 1D.

2 Réseaux de neurones bayésiens

2.1 Formalisme

Un réseau de neurones artificiels est composé de L couches, où la sortie de la $l^{\text{ième}}$ couche cachée, notée $f_l(\boldsymbol{x})$, est définie par :

$$f_l(\boldsymbol{x}) = \boldsymbol{W}^l \varphi(f_{l-1}(\boldsymbol{x})) + \boldsymbol{b}^l, \quad 1 \leq l \leq L-1, \tag{1}$$

où $\boldsymbol{x} \in \mathbb{R}^d$ est l'entrée de la première couche du réseau, φ la fonction d'activation (ici commune à toutes les couches cachées), $\boldsymbol{W}^l \in \mathbb{R}^{d_l \times d_{l-1}}$ la matrice des poids et $\boldsymbol{b}^l \in \mathbb{R}^{d_l}$ le vecteur des biais de la couche l. La sortie globale du réseau (sortie de la couche L) est donnée par :

$$f(\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b}) = \boldsymbol{W}^L \varphi(\cdots \varphi(\boldsymbol{W}^0 \boldsymbol{x} + \boldsymbol{b}^0)) + \boldsymbol{b}^L. \tag{2}$$

L'estimation des paramètres $(\boldsymbol{W}, \boldsymbol{b}) := \left\{\boldsymbol{W}^l, \boldsymbol{b}^l\right\}_{l=0}^{L}$ est effectuée en minimisant l'écart entre d'une part, les données disponibles $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i=1}^n$ où $\boldsymbol{y}_i$ est la $i^{\text{ème}}$ observation de la sortie du simulateur (ou du modèle) noté $\mathcal{M}$ tel que $\boldsymbol{y} = \mathcal{M}(\boldsymbol{x})$ et d'autre part, les prédictions du réseau $f(\boldsymbol{x}_i, \boldsymbol{W}, \boldsymbol{b})$.

Les réseaux de neurones artificiels sont performants pour apprendre des relations complexes, moyennant de disposer d'un nombre suffisant de données, mais ils ne fournissent pas directement une quantification de l'incertitude associée à chaque prédiction, ce qui en fait un frein à leur utilisation dans des domaines critiques tels que la sûreté nucléaire ou la conduite autonome. Les réseaux de neurones bayésiens (RNB) tentent de répondre à ce problème en combinant apprentissage profond et inférence bayésienne. Pour ce faire, le traitement bayésien repose sur trois étapes principales :

1. La spécification d'une distribution *a priori* $\pi(\boldsymbol{W}, \boldsymbol{b})$ pour représenter les croyances initiales sur les poids et les bais du réseau de neurones.

2. L'hypothèse d'une relation $\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b}) + \epsilon$ avec $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 \boldsymbol{I}_n)$ où σ_ϵ^2 représente la variance de l'erreur de modélisation par le réseau. La vraisemblance $\mathcal{L}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b})$ mesure alors l'adéquation entre les données $\boldsymbol{y}$ et les sorties $f(\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b})$ du réseau sachant les paramètres $(\boldsymbol{W}, \boldsymbol{b})$. A noter que σ_ϵ^2 est fixé dans un premier temps.

3. La mise à jour des croyances grâce à la formule de Bayes pour obtenir la distribution *a posteriori* :
$$\pi(\boldsymbol{W}, \boldsymbol{b} \mid \boldsymbol{x}, \boldsymbol{y}) \propto \mathcal{L}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b})\, \pi(\boldsymbol{W}, \boldsymbol{b}). \tag{3}$$

Les prédictions pour une nouvelle entrée $\boldsymbol{x}^\star$ s'obtiennent via la distribution prédictive :

$$\pi(\boldsymbol{y}^\star \mid \boldsymbol{x}^\star, \boldsymbol{x}, \boldsymbol{y}) = \int \mathcal{L}(\boldsymbol{y}^\star \mid \boldsymbol{x}^\star, \boldsymbol{W})\, \pi(\boldsymbol{W}, \boldsymbol{b} \mid \boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{W}\, d\boldsymbol{b}. \tag{4}$$

Cette intégrale ne peut généralement pas être calculée de manière analytique, des méthodes d'approximation de type MCMC comme la dynamique de Langevin stochastique sont utilisées pour échantillonner $\pi(\boldsymbol{y}^\star \mid \boldsymbol{x}^\star, \boldsymbol{x}, \boldsymbol{y})$ [Welling and Teh, 2011; Staber and Da Veiga, 2022].

2.2 Choix stratégique de la distribution *a priori*

Un aspect fondamental du cadre bayésien est le choix de $\pi(\boldsymbol{W}, \boldsymbol{b})$, qui influence significativement la qualité des prédictions du réseau. Dans ce travail, nous avons exploré différents *a priori*, en combinant certaines distributions existantes :

1. ***A priori* gaussien standard** sur les poids et les biais du réseau;
2. ***A priori* hiérarchique** sur les poids et les biais du réseau;
3. ***A priori horseshoe*** sur les poids des couches cachées du réseau afin de favoriser la parcimonie du réseau;
4. ***A priori* optimisé** via la minimisation de la distance de Wasserstein entre les trajectoires du RNB et celles d'une RPG [Tran et al., 2022].

Pour plus de détails sur ces différents *a priori* voir Fortuin [2022]; Tran et al. [2022]. Notons que l'*a priori horseshoe* est combiné ici avec un *a priori* hiérarchique qui est appliqué à tous les biais du réseau et aux poids de la couche de sortie. Sur l'exemple analytique 1D qui suit, les performances des différents *a priori* sont évaluées et comparées à celles d'une RPG avec une covariance stationnaire usuelle. A noter que la même covariance est utilisée pour l'*a priori* optimisé (approche 4).

3 Évaluation de l'impact de l'*a priori*

On considère la fonction analytique proposée par Sauer et al. [2023] et définie sur $[0, 1]$ par :

$$g(x) = \begin{cases} 1.35\cos(12\pi x) & \text{si } 0 \leq x \leq 0.33, \\ 1.35 & \text{si } 0.33 \leq x \leq 0.66, \\ 1.35\cos(6\pi x) & \text{si } 0.66 \leq x \leq 1. \end{cases}$$

Pour le RNB, on choisit une architecture suivante avec 3 couches cachées de 100 neurones chacune, une couche de sortie avec un seul neurone, et une fonction d'activation tangente hyperbolique dans les couches cachées. Et pour la RPG, on considère une moyenne constante et une fonction de covariance de type Matérn 5/2 [Rasmussen et al., 2006]. On considère un échantillon d'apprentissage i.i.d. $\{(x_i, y_i)\}_{i=1}^n$ de taille $n = 20$ où x_i est tiré uniformément sur $[0, 1]$ et $y_i = g(x_i)$. Les hyperparamètres de la RPG sont estimés par maximum de vraisemblance, tandis que l'estimation de la distribution *a posteriori* des poids et biais du RNB est réalisée à l'aide de la dynamique de Langevin stochastique. La variance de l'erreur de modélisation σ_ϵ^2 est fixée à 1% de la variance empirique de g calculée à partir de $\{y_i\}_{i=1}^n$. Une base de test indépendante de 100 points uniformément répartis sur $[0, 1]$ est ensuite utilisée pour comparer les performances des métamodèles RPG et RNB : les critères Q^2 et IAEα sont calculés permettant de quantifier respectivement la prédictivité (Q^2 le plus proche de 1) et la fiabilité des intervalles de crédibilité (IAEα le plus proche de 0). Voir [Marrel and Iooss, 2024] pour plus de détails sur ces critères. .

La Figure 1 illustre les résultats obtenus sur un échantillon d'apprentissage (de taille $n = 20$). Les valeurs correspondantes des Q^2 et IAE_α sont données par le Tableau 1. On observe tout d'abord que la médiane des RNB avec un *a priori* gaussien ou hiérarchique conduit à une très bonne prédictivité, similaire à celle de la RPG ($Q^2 \geq 0.95$). Seul le RNB avec un *prior horseshoe* présente un performance très inférieure (Q^2 à 0.66) : cela pourrait s'expliquer par la trop grande parcimonie imposée par l'*a priori* dont l'influence prédomine, compte tenu de la faible valeur de n. Concernant la qualité des intervalles de crédibilité, ceux fournis par la RPG sont trop conservatifs, en particulier sur la partie constante de la fonction cible ($x \in [0.33, 0.66]$), tandis que ceux des RNB sont plus réguliers et cohérents avec la forme de la fonction. Cette observation est confirmée par le calcul des IAE_α: les RNB permettent de considérablement réduire sa valeur qui passe de 0.40 pour la RPG à moins de 0.25 avec la plupart des RNB. Sur cet exemple, le meilleur compromis entre prédictivité et qualité des intervalles est ainsi obtenu avec les RNB avec *a priori* gaussiens (standard et optimisé).

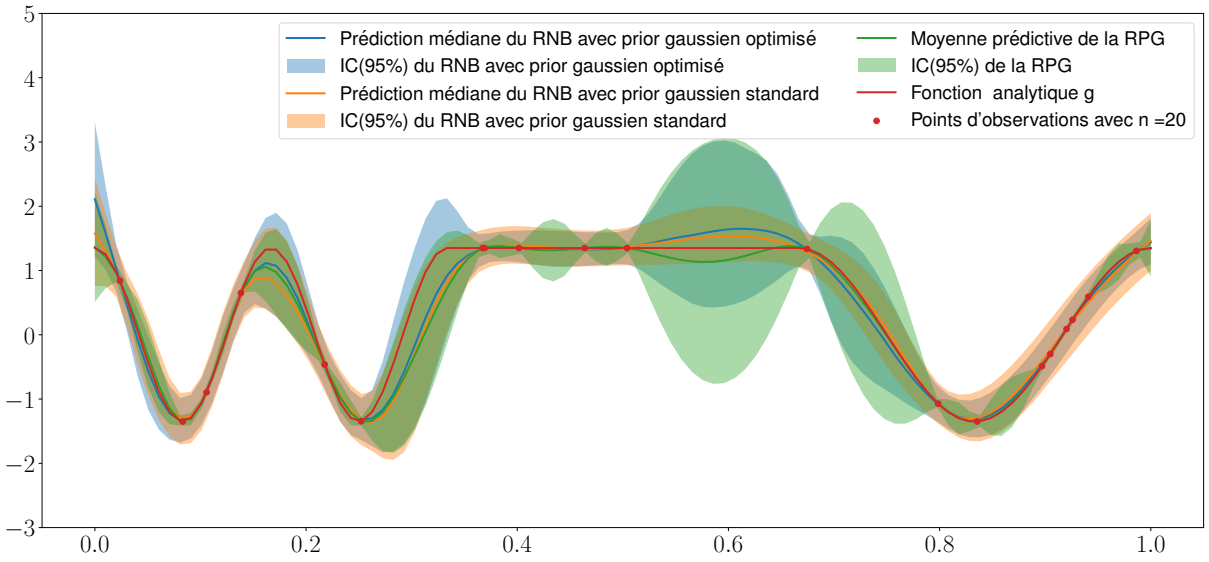


Figure 1: Prédictions (moyenne prédictive pour la RPG et prédiction médiane pour les RNB) et intervalles de crédibilité à 95%.

Métamodèle	Q^2	IAE_α
RPG (Covariance Matérn 5/2)	**0.95**	0.4
RNB avec *a priori* gaussien standard	**0.95**	**0.16**
RNB avec *a priori* gaussien optimisé	**0.96**	0.20
RNB avec *a priori* hiérarchique	**0.97**	0.22
RNB avec *a priori* hiérarchique optimisé	**0.96**	0.24
RNB avec *a priori horseshoe* + hiérarchique	0.66	0.20

Table 1: Comparaison des performances des RPG et RNB en termes de Q^2 et IAE_α.

Conclusion

Ces premiers résultats suggèrent que l'optimisation des distributions *a priori* des paramètres des RNB pourrait jouer un rôle important dans l'amélioration des performances de ces métamodèles, et ouvrir la voie à leur utilisation pour la quantification des incertitudes en simulation numérique. Sur l'exemple considéré, cette optimisation permet d'obtenir des RNB avec une prédictivité aussi bonne que la RPG (considérée aujourd'hui comme le métamodèle de référence) mais avec des intervalles de crédibilité capables de mieux s'adapter à la forme de la fonction cible. Ces premiers résultats seront à confirmer sur d'autres échantillons de données d'apprentissage, et en considérant d'autres fonctions analytiques de dimension variée. Il sera aussi pertinent de considérer des architectures de réseau plus complexe, combiné à un plus grand nombre de données d'apprentissage, ou encore une estimation d'une distribution *a posteriori* pour σ_ϵ^2. Celle-ci pourrait permettre d'augmenter la capacité des RNB à capturer la complexité de la fonction (ou du modèle) à émuler.

Bibliographie

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991). Bayesian Prediction of Deterministic Functions, with Applications to the Design and Analysis of Computer Experiments. Journal of the American Statistical Association, 86(416):953–963.

Fortuin, V. (2022). Priors in Bayesian Deep Learning: A review. International Statistical Review, 90(3):563–591.

Marrel, A. and Iooss, B. (2024). Probabilistic surrogate modeling by Gaussian process: A review on recent insights in estimation and validation. Reliability Engineering & System Safety, page 110094.

Rasmussen, C. E., Williams, C. K., et al. (2006). Gaussian Processes for Machine Learning, volume 1. Springer.

Sauer, A., Gramacy, R. B., and Higdon, D. (2023). Active learning for Deep Gaussian process surrogates. Technometrics, 65(1):4–18.

Staber, B. and Da Veiga, S. (2022). Benchmarking Bayesian neural networks and evaluation metrics for regression tasks. arXiv preprint arXiv:2206.06779.

Tran, B.-H., Rossi, S., Milios, D., and Filippone, M. (2022). All You Need is a Good Functional Prior for Bayesian Deep Learning. Journal of Machine Learning Research, 23(74):1–56.

Welling, M. and Teh, Y. W. (2011). Bayesian Learning via Stochastic Gradient Langevin Dynamics. In Proceedings of the 28th international conference on machine learning (ICML-11), pages 681–688. Citeseer.

Dirichlet-Based Monte Carlo Dropout for Uncertainty Estimation in Neural Networks

Rouaa Hoblos & Noura Dridi & Noureddine Zerhouni & Zeina Al Masry

Université Marie et Louis Pasteur, SUPMICROTECH, CNRS, Institut FEMTO-ST, F-25000 Besançon, France
rouaa.hoblos@femto-st.fr, noura.dridi@femto-st.fr, noureddine.zerhouni@femto-st.fr, zeina.almasry@femto-st.fr

Résumé. Les réseaux neuronaux traditionnels fournissent des prédictions déterministes sans estimation explicite de l'incertitude. Bien que les réseaux neuronaux bayésiens (BNNs) offrent une approche théoriquement fondée pour la quantification d'incertitude, leur complexité computationnelle limite leur passage à l'échelle. Le Dropout de Monte Carlo (MC Dropout), initialement introduit comme une technique de régularisation, a été démontré comme une approximation de l'inférence bayésienne en permettant une modélisation probabiliste via plusieurs passes stochastiques. Dans ce travail, nous renforçons l'estimation de l'incertitude en intégrant un cadre basé sur la distribution de Dirichlet au sein de MC Dropout. Plus précisément, nous exploitons la formulation proposée dans Sensoy et al. (2018), où les probabilités de classe sont modélisées par une distribution de Dirichlet, permettant une représentation plus informative de l'incertitude. Notre approche conserve l'efficacité computationnelle du MC Dropout tout en améliorant la qualité des estimations d'incertitude. Nous discutons les fondements théoriques de notre méthode et la comparons aux techniques existantes de quantification d'incertitude. Nos résultats mettent en évidence l'efficacité de la méthode proposée pour produire des estimations d'incertitude bien calibrées, offrant une solution pratique pour les modèles d'apprentissage profond intégrant l'incertitude.

Mots-clés. Quantification d'incertitude, Monte Carlo Dropout, Réseau de Neurones, Loi des Grands Nombres

Abstract. Traditional neural networks provide deterministic predictions without inherent uncertainty estimates. While Bayesian Neural Networks (BNNs) offer a principled approach to uncertainty quantification, their computational complexity limits scalability. Monte Carlo (MC) Dropout, initially introduced as a regularization technique, has been shown to approximate Bayesian inference by enabling probabilistic modeling through multiple stochastic forward passes. In this work, we enhance uncertainty estimation in deep learning by integrating a Dirichlet-based framework within MC Dropout. Specifically, we leverage the formulation proposed by Sensoy et al. (2018), where class probabilities are modeled using a Dirichlet distribution, allowing for a more informative uncertainty representation. The proposed approach maintains the computational efficiency of MC Dropout while improving the quality of uncertainty estimates. We discuss the theoretical foundations of our method and compare it with existing uncertainty quantification techniques. The results highlight the effectiveness of the proposed method in producing well-calibrated uncertainty estimates, offering a practical solution for uncertainty-aware deep learning models.

Keywords. Uncertainty quantification, Monte Carlo dropout, Neural Network, Law of Large Numbers

1 Introduction

Effective decision-making relies on the ability to make precise and reliable predictions including quantifying uncertainty associated with these predictions, this means the confidence level to attribute to these predictions. This is particularly relevant for critical domains such as medical field or autonomous vehicle control. Several approaches have been developed, including Bayesian Neural Networks (BNNs) [1], ensemble methods [2], and single deterministic models [3]. BNNs estimate uncertainty by modeling neural network weights as probability distributions. A variational interpretation of the BNN is proposed in [4]. The main result is that optimizing the loss function of a NN with dropout is equivalent to a Bayesian variational approximation of a Gaussian process. This method enables probabilistic modeling by leveraging dropout at both training and inference stages, allowing multiple stochastic forward passes to estimate uncertainty. The method is computationally efficient compared to classical BNNs [1], which are computationally expensive for posterior inference. On the other hand, ensemble methods train multiple models and quantify uncertainty using the variability among their predictions. Single deterministic methods, such as evidential deep learning [5], use a Dirichlet distribution to represent the class probability. The parameters of this distribution are related to the prediction and uncertainty. In [5], the authors propose a new method called Deep Evidential Classification (DEC), an evidential neural network to learn these parameters in classification tasks. In [6], the author proposes an Information Aware Dirichlet (IAD) network to learn a Dirichlet prior distribution on predictive distributions. In this work, we propose to use the same hypothesis for the class probability, however, we propose a new method to calculate Dirichlet parameters and therefore the uncertainty. This method leverages MC Dropout while maintaining computational efficiency and improving uncertainty estimation through a theoretically grounded probabilistic formulation. The performance of the method is confirmed on tabular and image datasets and compared with the approach proposed in [5] and [6] using different metrics.

2 Proposed approach

Let $\{(X_i, Y_i)\}_{i=1}^N$ be a set of data samples, where X_i represents an input (e.g., an image or a feature vector), and Y_i is the corresponding label in a classification task, with $Y_i \in \mathcal{Y} = \{1, 2, \ldots, K\}$. A neural network parameterized by θ is used to predict the class probability distribution $p_\theta(X_i) = (p_1, p_2, \ldots, p_K)$, where p_k is the predicted probability for class k, obtained by applying the softmax function to the network's logits.

Uncertainty and Evidence. In [5], Sensoy has proposed calculating the uncertainty using evidence learned by the neural network. The NN outputs evidence which is used in the calculation of the parameters of a Dirichlet distribution. These parameters are used to calculate the predicted probabilities and the uncertainties associated to them. The uncertainty is given

by:

$$u = \frac{K}{S} \tag{1}$$

with $S = \sum_{k=0}^{K}(e_k + 1)$, with e_k the evidence derived for the kth singleton. The authors [5] propose to replace the softmax activation function, classically used for classification, by the $ReLU$ function to obtain positive values corresponding to the evidence. In [6], Tsiligkaridis uses the softplus activation function. The predicted probabilities are the expected value calculated using the properties of a Dirichlet distribution: $p_k = \frac{\alpha_k}{S}$ where p_k is the probability associated to the k^{th} class.

Dirichlet Based Uncertainty Estimation-MC Dropout (DBUE-Dropout). As proposed in [5], the class probability vector $\mathbf{p} = (p_1, p_2, \ldots, p_K)$ is assumed to follow a *Dirichlet distribution* parameterized by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$:

$$p(\mathbf{p}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \tag{2}$$

where $B(\boldsymbol{\alpha})$ is the multivariate Beta function.

The variance of the Dirichlet distribution can be written as [7]:

$$var(p_k) = \frac{\mathbb{E}[p_k](1 - \mathbb{E}[p_k])}{1 + \sum_k \alpha_k}. \tag{3}$$

Therefore, S would be given by:

$$S = \sum_k \alpha_k = \frac{\mathbb{E}[p_k](1 - \mathbb{E}[p_k])}{\mathrm{Var}(p_k)} - 1 \tag{4}$$

where $\mathbb{E}[p_k]$ is the expected probability for class k and $\mathrm{Var}(p_k)$ is the variance associated to this predicted probability.

In the context of classification, a NN architecture is designed to output the class probabilities. Besides, by activating the dropout on the test, we obtain a different probability for each new forward. It is worth pointing out that MC dropout computes the standard deviation to estimate uncertainty; however, the standard deviation measures the variation but doesn't convey the full characteristics of a distribution. We propose to represent the probability of classes by the Dirichlet distribution. Indeed, a classification problem with K labels is a multinomial opinion equivalent to a Dirichlet probability distribution function[8]. Given an input X_i, we perform L stochastic forward passes with dropout enabled, obtaining L probability vectors: $p^{(1)}, p^{(2)}, \ldots, p^{(L)}$. For each class k, the expected probability and variance are computed empirically. Besides, we estimate $\mathbb{E}[p_k]$ and $\mathrm{Var}(p_k)$ using the Law of Large Numbers (LLN) [9]. The LLN states that the empirical mean of a sequence of independent and identically distributed (i.i.d.) random variables converges to the true expectation as the number of samples increases. Despite its simplicity, this process makes our proposed method a fully coherent Bayesian method with a connection to evidence. Using the empirical estimates, S is computed using (4) and used to obtain the final uncertainty measure as in (1).

A higher S means more evidence and less uncertainty and vice versa. It is important to note that the uncertainty is calculated using the absolute value of S to ensure it is non-negative.

Algorithm. An algorithm of the proposed methodology is shown below (Algorithm 1).

Algorithm 1 Dirichlet Based Uncertainty Estimation-MC Dropout

Require: Dataset $\{(X_i, Y_i)\}_{i=1}^{n}$, trained neural network $p_\theta(X)$, number of MC Dropout iterations L, number of classes K

Ensure: Uncertainty estimates u_i for each sample X_i

Initialize: Neural network $p_\theta(X)$ with MC Dropout enabled

for each sample X_i in the test set **do**

Collect L stochastic forward passes:

$p_i^{(l)} = p_\theta(X_i)$ for $l = 1, \ldots, L$

Determine the predicted class:

$c_i = \arg\max_k \left(\frac{1}{L} \sum_{l=1}^{L} p_i^{(l)}(k) \right)$

Extract probabilities of the predicted class:

$p_i^{(l)}(c_i)$ for $l = 1, \ldots, L$

Compute the empirical expectation for predicted class:

$\mathbb{E}[p_i(c_i)] = \frac{1}{L} \sum_{l=1}^{L} p_i^{(l)}(c_i)$

Compute the empirical variance:

$\mathrm{Var}(p_i(c_i)) = \frac{1}{L-1} \sum_{l=1}^{L} \left(p_i^{(l)}(c_i) - \mathbb{E}[p_i(c_i)] \right)^2$

Compute Dirichlet strength parameter for predicted class:

$S_i = \frac{\mathbb{E}[p_i(c_i)](1-\mathbb{E}[p_i(c_i)])}{\mathrm{Var}(p_i(c_i))} - 1$

Compute uncertainty:

$u_i = \frac{K}{S_i}$

end for

Return: Uncertainty estimates $\{u_i\}_{i=1}^{n}$

3 Results

The proposed method is evaluated on both tabular and image datasets, specifically MNIST and Titanic. To assess the effectiveness of the UQ method in detecting data shifts, we introduce noisy data and Out-of-Distribution (OOD) datasets. For MNIST, Gaussian noise with a mean of 2 and standard deviation of 100 is added to test images, while for Titanic, noise is selectively applied to different features with a mean of either 0 or 1 and a maximum standard deviation of 3. The Fashion-MNIST (FMNIST) dataset serves as an OOD dataset for MNIST, whereas the Forest Fires dataset is used for Titanic. Training is conducted with 350 epochs and 100 Monte Carlo dropout samples ($p_{drop} = 0.25$) for Titanic, while MNIST is trained for 150 epochs with 100 dropout samples ($p_{drop} = 0.4$). Furthermore the method is compared to DEC [5] and IAD [6]. For the DEC, the NN outputs the evidence calculated using the ReLU activation function, and the uncertainty is calculated using (1). As for the

IAD, the NN outputs the evidence calculated using the Softplus activation function, and the uncertainty is calculated using (1).

Table 1: Performance Metrics Across Titanic Dataset with Different Methods.

Method	Accuracy				Uncertainty			
	Train	Test	Noisy	OOD	Train	Test	Noisy	OOD
IAD [6]	86.82	76.92	**72.032**	**0.00**	0.53	0.54	0.51	0.012
DEC [5]	88.22	**80.42**	65.73	**0.00**	0.59	0.57	0.48	0.0033
DBUE-Dropout	**89.6**	77.62	62.94	0.19	**0.027**	**0.026**	**0.028**	**2.0**

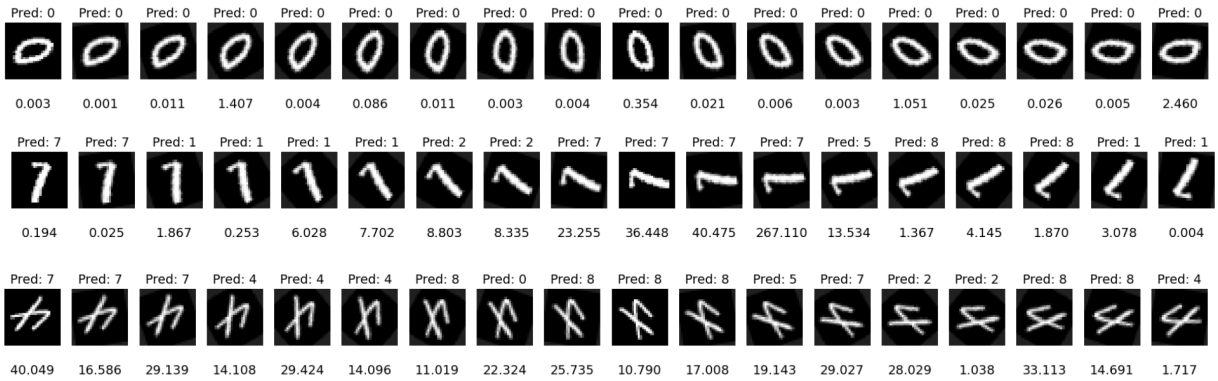


Figure 1: Rotated images with their associated uncertainties.

Table 1 illustrates the classification performance as well as the UQ performance using the three methods on Titanic dataset. The uncertainty is calculated per sample and the median over train, test and noisy sets are given (to avoid effect of outliers). It is noted that DEC and IAD are unable to detect OOD data since the uncertainty decreases while the accuracy is decreasing meaning the algorithm is confident in its wrong decisions. Meanwhile, DBUE-Dropout can efficiently detect the noisy and OOD data. This means accuracy decreases on OOD data and uncertainty increases on noisy and OOD data. As for images, the DBUE-Dropout method performs well on the MNIST dataset. The method demonstrates high performance with a training accuracy of 99.95%, a test accuracy of 99.25%, a noisy test accuracy of 9.74%, and an OOD accuracy of 6.78%. As for the uncertainty measure, it is 0.27 for training, 0.33 for testing, 0.53 for noisy inputs, and 19.19 for OOD data. The stranger the image becomes to the trained model, the higher the uncertainty. This can be shown by the increase in the median of uncertainty across different datasets.

To assess DBUE-Dropout's performance on images, some were rotated (0–360°), and their uncertainties were analyzed (Figure 1). The digit 0 maintains low uncertainty due to its rotational invariance. As for the second image, uncertainty spikes at ambiguous angles, especially when horizontal, but remains low when clearly resembling a 1 or 7. The third row represents an image of digit 4 highlights that even correct predictions can exhibit high uncertainty. For instance, when inverted, uncertainty remains elevated due to visual ambiguity. However, as

the digit returns to an upright position, uncertainty drops from almost 14 to 2, reinforcing the model's confidence in more recognizable orientations.

4 Conclusion

A new method for uncertainty quantification in classification tasks is proposed. The idea is to construct a deep evidential neural network with MC dropout. Given different samples of the class probability, the parameters of the Dirichlet distribution are calculated and used to evaluate the uncertainty. Results confirm the overall performance in terms of classification and uncertainty quantification. Besides, comparison with state-of-the-art methods illustrates the potential of the proposed method to detect distribution shift through noisy and out-of-distribution (OOD) data across both tabular and image datasets. Future work will focus on enhancing the method to improve performance on noisy data and generalize to imbalanced datasets.

References

[1] Radford M. Neal. Bayesian learning for neural networks. 1995.

[2] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[3] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[4] Gal.Y and Ghahramani.Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. 2016.

[5] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[6] Theodoros Tsiligkaridis. Information aware max-norm dirichlet networks for predictive uncertainty estimation. *Neural Networks*, 135:105–114, 2021.

[7] Thomas Minka. Estimating a dirichlet distribution, 2000.

[8] Audun Jsang. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018.

[9] Michel Loève and M Loève. *Elementary probability theory*. Springer, 1977.

Régression Logistique Optimale pour la Classification de Distribution Induite par un Arbre

Baptiste Schall [1] & Rodolphe Anty [2] & Lionel Fillatre [3]

[1] *Université Côte d'Azur, France, bschall@i3s.unice.fr*
[2] *CHU Nice, Unité d'hépatologie, France, anty.r@chu-nice.fr*
[3] *Université Côte d'Azur, France, lionel.fillatre@i3s.unice.fr*

Résumé

L'importance des interactions inter-caractéristiques pour les tâches de classification est prouvée par l'abondance de méthodes de Machine Learning très performantes les utilisant. Dans cet article, nous nous inspirons des Réseaux Bayésiens (RB), une méthode de classification connue pour sa représentation graphique des liens entre caractéristiques. Dans le cas discret, nous supposons que les caractéristiques suivent une distribution induite par un arbre pour chaque classe. Ensuite, nous montrons que le classificateur de Bayes optimal coïncide avec une régression logistique dont les caractéristiques sont encodées avec un schéma spécifique. Nous démontrons les avantages d'une telle approche à la fois théoriquement et numériquement, notamment en montrant que l'étape d'apprentissage converge toujours vers une solution unique. Des expériences sur données simulées confirment l'efficacité du modèle présenté.

Mots-clés. Classification binaire, Régression Logistique, Réseaux Bayésiens, Classifieur de Bayes, Encodage One-hot

Abstract

The importance of features interactions for classification tasks between two classes is proven by the abundance of high-performance machine learning methods using them. In this paper, we take our inspiration from Bayesian Networks (BNs), a famous classification method whose main asset is the graphical representation of links between features. In the case of discrete feature, we assume that the features follow a tree-based distribution under each class. Then, we show that the optimal Bayes classifier coincides with a logistic regression whose features are one-hot encoded with a specific scheme. We show the advantages of such an approach both theoretically and numerically, especially showing that the learning step always converges toward a unique solution. Simulated experiments confirm the efficiency of the one-hot logistic regression with feature interaction encoding.

Keywords. Binary Binary Classification, Logistic Regression, Bayesian Network, Bayes Classifier, One-hot encoding

1 Introduction

Les modèles de Machine Learning (ML) les plus performants aujourd'hui sont des modèles non linéaires qui tiennent compte des interactions entre les caractéristiques, tels que les algorithmes de boosting. L'inconvénient de ces algorithmes est leur nature de "boîte noire" qui rend l'interprétation de leurs règles de décision fastidieuse. Nous nous focaliserons des modèles intrinsèquement interprétables, tels que les réseaux bayésiens (BN) [1], combinés à un algorithme de classification simple pour leur apprentissage, la régression logistique (LR).

La principale faiblesse des BN réside dans le fait que l'apprentissage d'un modèle graphique, est NP-complet [2]. Une des manières de contourner ce problème est : l'apprentissage de graphe d'interaction contraint. Utiliser un modèle graphique contraint tel qu'un arbre peut déjà fournir de meilleurs résultats en terme de classification que les modèles naïfs. Dans [3] est présenté l'algorithme d'apprentissage d'arbre de Chow et Liu, bien qu'optimal au sens de l'estimation, il est remis en question quant à sa pertinence pour les tâches de classification [4].

Dans cet article, nous développons une nouvelle LR avec des caractéristiques one-hot encodées, équivalente au classifieur de Bayes optimal discriminant deux distributions induites par des arbres. Cette nouvelle LR est une extension du classifieur naïf bayésien discret, qui omet les interactions entre les caractéristiques [5]. Nous établissons d'abord théoriquement l'équivalence entre notre LR et le classifieur de Bayes optimal pour des distributions induites par un arbre. Notre LR encode les valeurs marginales et marginales par paires des caractéristiques. Ensuite, nous étudions la convexité de la méthode proposée en montrant qu'une réduction du nombre de paramètres est nécessaire pour garantir la stricte convexité du critère. Enfin, nous présentons des expériences sur des jeux de données simulés, démontrant l'importance des interactions entre les caractéristiques ainsi que la convergence de notre méthode. Nous étudions également la robustesse de notre méthode en haute dimension.

La section 2 montre que le classifieur de Bayes optimal et la LRT coïncident à condition que les caractéristiques soient correctement encodées. La section 3 introduit l'encodage suffisant garantissant la convergence de l'algorithme d'apprentissage vers un unique modèle. La section 4 illustres la pertinence de notre méthode avec des données simulées en montrant sa consistance sur différent type de données. La section 5 conclue l'article.

2 Encodage de distribution induite par des arbres

Nous considérons une base de données $\mathcal{D} = \left\{(\underline{x}^{(i)}, c^{(i)})\right\}_{i=1}^{N}$ composée de N échantillons $x^{(i)} \in \mathbb{R}^d$ avec leur label $c^{(i)} \in 0, 1$. Chaque composante x_i du vecteur $\underline{x} = (x_1, \ldots, x_d)$ appartient à un ensemble $\Omega_i \subset \mathbb{R}$. Le vecteur $\underline{x}$, respectivement le label c, est une réalisation du vecteur aléatoire $\underline{X} = (X_1, \ldots, X_d)$, respectivement de la variable aléatoire binaire C. Nous étudierons uniquement la classification binaire entre deux distributions p et q, bien qu'une extension au multi-classes soit possible. Notre objectif est de montrer que le classifieur de Bayes optimal coïncide avec une régression logistique lorsque nous utilisons un encodage one-hot approprié. Pour ce faire, nous devons travailler avec des variables discrètes. Par conséquent, nous supposons que chaque variable x_i peut prendre t valeurs réelles $\omega_{i,1}$, ...,

$\omega_{i,t}$. Lorsque x_i est continue, nous pouvons quantifier la variable pour obtenir les valeurs $\omega_{i,j}$. Nous adoptons la notation $p(\underline{x}) = \mathbb{P}_{X|C}(\underline{x} \mid C = 0)$ et $q(\underline{x}) = \mathbb{P}_{X|C}(\underline{x} \mid C = 1)$. Nous supposons que p et q sont modélisés par une distribution de probabilités induite d'un arbre. Un arbre est défini de manière unique par un d-uplet $\underline{m}^p = (m_1^p, \ldots, m_d^p)$, où le i-ième élément de $\underline{m}^p$ indique le parent unique $x_{m_i^p}$ de la variable x_i. Nous supposons que $x_1, \ldots, x_d$ sont ordonnés de manière à ce que $m_i^p < i$ dans (2). Par convention, x_1 est la racine de l'arbre : elle n'a pas de parent ($m_1^p = 0$). La distribution p (et également q) se factorise de la sorte :

$$p(\underline{x}) = \prod_{i=1}^{d} p_i(x_i) p'_i(x_i, x_{m_i^p}), \quad \text{et} \quad p'_i(x_i, x_{m_i^p}) = \frac{p_{i,m_i^p}(x_i, x_{m_i^p})}{p_i(x_i) p_{m_i^p}(x_{m_i^p})}, \quad p'_1(x_1, x_{m_1}^p) = 1, \tag{1}$$

où $p_i(x_i)$ est la marginale de la variable aléatoire x_i, et $p_{i,j}(x_i, x_j)$ est la marginale conjointe du couple (x_i, x_j). En raison de l'hypothèse de structure en arbre, $\mathcal{E}^p$ l'ensemble des arêtes associé à la distribution p a la forme suivante [4] :

$$\mathcal{E}^p = \{(i, m_i^p) : 1 \leq i \leq d, m_1^p = 0, 1 \leq m_i^p < i\}. \tag{2}$$

L'apprentissage d'une distribution structurée en arbre consiste à apprendre à la fois $\underline{m}^p$, $p_i(x_i)$ et $p'_i(x_i, x_{m_i}^p)$. Une fois la distribution structurée en arbre apprise, le classifieur de Bayes optimal $f^*(\underline{x})$ est donné par [6] :

$$f^*_{p,q}(\underline{x}) = \begin{cases} 1 \text{ if } \varphi^*_{p,q}(\underline{x}) > 0, \\ 0 \text{ sinon,} \end{cases} \quad \text{et} \quad \varphi^*_{p,q}(\underline{x}) = \alpha_0 + \sum_{i=1}^{d} \underline{\alpha}_i^T \underline{\tilde{x}}_i + \sum_{i=1}^{d} \sum_{k \in \{p,q\}} \underline{\alpha}_{i,k}^T \left(\underline{\tilde{x}}_i \otimes \underline{\tilde{x}}_{m_i^k} \right), \tag{3}$$

où on introduit des variables indicatrices $\underline{\tilde{x}}_i = \left[\tilde{x}_{i,1}, \ldots, \tilde{x}_{i,t}\right]$ telles que $\tilde{x}_{i,j} = 1$ pour les observations telles que $x_i = \omega_{i,j}$, et $\tilde{x}_{i,j} = 0$ sinon. Le produit de Kronecker est noté $\otimes$ et $\underline{x}^T$ est la transposée de $\underline{x}$. Les vecteurs de coefficients dans $\varphi^*_{p,q}(\underline{x})$ sont donnés par

$$\underline{\alpha}_i = \left[\alpha_{i,1}, \ldots, \alpha_{i,t}\right], \quad \underline{\tilde{x}}_i = \left[\tilde{x}_{i,1}, \ldots, \tilde{x}_{i,t}\right], \quad \underline{\alpha}_{i,k} = \left[\alpha_{i,1,m_i^k,1}, \ldots, \alpha_{i,1,m_i^k,t}, \ldots, \alpha_{i,t,m_i^k,t}\right] \in \mathbb{R}^{t^2},$$

tels que α_0 et $\alpha_{i,j,m_i^k,\ell}$ sont définis par

$$\alpha_0 = \ln \frac{\mathbb{P}_C(C=1)}{\mathbb{P}_C(C=0)}, \; \alpha_{i,j} = \ln \frac{q_i(\omega_{i,j})}{p_i(\omega_{i,j})}, \; \alpha_{i,j,m_i^k,\ell} = \begin{cases} -\ln p'_i(\omega_{i,j}, \omega_{m_i^p,\ell}) & \text{if } (i, m_i^p) \in \mathcal{E}^p, \\ \ln q'_i(\omega_{i,j}, \omega_{m_i^q,\ell}) & \text{if } (i, m_i^q) \in \mathcal{E}^q. \end{cases} \tag{4}$$

La fonction de score dans (3) est la somme d'une constante, d'un score marginal basé sur les caractéristiques $\tilde{x}_{i,j}$ et d'un score marginal par paire basé sur les croisements de caractéristiques $\tilde{x}_{i,j,k,\ell} = \tilde{x}_{i,j}\tilde{x}_{k,\ell}$. Les croisements de caractéristiques sont déduits des arêtes de l'arbre. Cette fonction de score contient $m = 1 + dt + et^2$ paramètres, où e est le nombre d'arêtes dans $\mathcal{E}^p \cup \mathcal{E}^q$. Si nous concaténons tous les paramètres dans un seul vecteur $\underline{\alpha}$ (y compris la constante α_0) et toutes les caractéristiques et croisements de caractéristiques dans un seul vecteur $\underline{\tilde{x}}$ (incluant la valeur constante 1 comme premier élément), nous obtenons la régression logistique suivante, appelée LRT,

$$\varphi^*_{p,q}(\underline{x}) = \underline{\alpha}^T \underline{\tilde{x}}. \tag{5}$$

Le modèle LRT peut représenter le classifieur de Bayes optimal entre n'importe quelles paires de distributions arborées, à condition d'utiliser un encodage adéquat.

3 Estimation de la LRT et Encodage suffisant

La LRT est estimée avec l'entropie croisée binaire $\mathcal{L}(\underline{\alpha})$. Soit $\hat{\underline{\alpha}}$ un estimateur qui minimise $\mathcal{L}(\underline{\alpha})$. Nous nous attendons bien sûr à ce que $\hat{\underline{\alpha}}$ soit proche du vecteur inconnu $\underline{\alpha}$. La LRT (5) est entièrement caractérisée par l'ensemble fini des scores $\sigma(\hat{\underline{\alpha}}^\top \tilde{\underline{x}}^{(i)})$ calculés sur le jeu de données $\mathcal{D}$. Ainsi, la LRT est uniquement définie par le vecteur de scores $\tilde{X}\hat{\underline{\alpha}}$ où $\tilde{X}$ est la matrice des données d'entrée (une ligne par échantillon) : $\tilde{X} = \left[\tilde{\underline{x}}^{(i)\top}, \ldots, \tilde{\underline{x}}^{(N)\top}\right]^\top \in \mathbb{R}^{N\times m}$. Le hessien $\nabla^2\mathcal{L}(\underline{\alpha})$ de $\mathcal{L}(\underline{\alpha})$ peut s'écrire :

$$\nabla^2\mathcal{L}(\underline{\alpha}) = \tilde{X}^\top D(\underline{\alpha})\tilde{X}, \qquad D(\underline{\alpha}) = \operatorname{diag}\left(\left[\sigma(\underline{\alpha}^\top\tilde{\underline{x}}^{(i)})\left(1-\sigma(\underline{\alpha}^\top\tilde{\underline{x}}^{(i)})\right)\right]\right). \tag{6}$$

Comme $D(\underline{\alpha})$ est diagonale avec tous ses éléments diagonaux non nuls, le rang de $\nabla^2\mathcal{L}(\underline{\alpha})$ est égal à celui de la matrice de données $\tilde{X}$. Nous en déduisons le lemme suivant.

Lemma 1. *Le hessien $\nabla^2\mathcal{L}(\underline{\alpha})$ est semi-défini positif de rang $1 + d(t-1) + e(t-1)^2 < m$. La fonction $\mathcal{L}(\underline{\alpha})$ n'est pas strictement convexe et l'estimateur $\hat{\underline{\alpha}}$, s'il existe, n'est pas unique.*

Nous n'avons donc aucune garantie concernant la fiabilité de l'estimation $\hat{\underline{\alpha}}$, celle ci n'étant pas unique. Afin de garantir l'unicité de l'estimation, la proposition suivante introduit un nouvel encodage, appelé encodage suffisant, qui garantit la convexité stricte de $\mathcal{L}(\underline{\alpha})$.

Proposition 1. *Let $r = 1 + d(t-1) + e(t-1)^2$. Il existe une matrice de permutation Q telle que $\tilde{X}$ et $\underline{\alpha}$ satisfont :*

$$\tilde{X}Q = \left[\tilde{S} \;\middle|\; \tilde{R}\right], \quad Q^\top\underline{\alpha} = \begin{bmatrix}\underline{\theta}^S \\ \underline{\theta}^R\end{bmatrix}, \tag{7}$$

où $\tilde{S}$ est de taille $N \times r$, $\tilde{R}$ est de taille $N \times (m-r)$, $\underline{\theta}^S \in \mathbb{R}^r$ et $\underline{\theta}^R \in \mathbb{R}^{m-r}$. Les propriétés suivantes sont satisfaites :

- *$\tilde{S}$ est une matrice de rang plein colonne r,*
- *Il existe une matrice L de taille $r \times (m-r)$ telle que $\tilde{R} = \tilde{S}L$.*
- *Les scores $\tilde{X}\underline{\alpha}$ sur $\mathcal{D}$ sont préservés i.e*

$$\tilde{X}\underline{\alpha} = \tilde{S}(\underline{\theta}^S + L\underline{\theta}^R) = \tilde{S}\underline{\theta}, \;\; \underline{\theta} = \underline{\theta}^S + L\underline{\theta}^R \tag{8}$$

La Proposition 1 introduit une nouvelle matrice de codage $\tilde{S}$ déduite du codage initial $\tilde{X}$. Ce nouveau codage induit une forme alternative de $\varphi^*_{p,q}(\underline{x})$ avec moins de paramètres

$$\varphi^*_{p,q}(\underline{x}) = \underline{\theta}^T\underline{\tilde{s}}, \tag{9}$$

ou $\underline{\tilde{s}}$ est l'encodage suffisant de $\underline{\tilde{x}}$ déduit de la factorisation (7).

Lemma 2. *Soit $\mathcal{L}(\underline{\theta})$ la fonction de perte à minimiser pour estimer les paramètres $\underline{\theta}$ dans (9). Le hessien $\nabla^2\mathcal{L}(\underline{\theta}) \in \mathbb{R}^{r\times r}$ de $\mathcal{L}(\underline{\theta})$ est défini positif et de rang r. La fonction $\mathcal{L}(\underline{\theta})$ est donc strictement convexe, et l'estimateur $\hat{\underline{\theta}}$ qui minimise $\mathcal{L}(\underline{\theta})$, s'il existe, est unique.*

D'après le Lemme 2, nous avons la garantie d'obtenir une estimation unique $\hat{\underline{\theta}}$.

4 Expériences

Cette section a pour but de montrer les performances de notre classifieur. Nous comparons LRT avec des méthodes telles que LR, LR avec encodage one-hot (LRO) [5], ainsi que deux classifieurs basés sur des réseaux bayésiens. Le premier utilise deux structures graphiques correspondant à des arbres de Chow & Liu [3], spécifiques à chaque valeur de classe, et est appelé CL. Le second correspond à un TAN. Un TAN est une extension du classifieur naïf bayésien permettant au maximum une connexion entre les caractéristiques d'entrée, en plus de la connexion de chaque caractéristique à la classe. : $p \sim \mathcal{N}(0, \Sigma_p)$ et $q \sim \mathcal{N}(0, \Sigma_q)$. Nous construisons deux ensembles de données : le premier avec $d = 3$ et une matrice de corrélation tirée de [4] pour observer le comportement des méthodes en faible dimension, et le second avec $d = 20$. Ces ensembles de données sont respectivement appelés *GGM3D* et *GGM20D* (GGM : Graphical Gaussian Models).

La simulation de notre propre ensemble de données nous permet de calculer le classifieur théorique pour les méthodes LRT et TAN (notés avec le suffixe th.). La première expérience illustre la contribution des interactions entre caractéristiques. La Fig. 1 montre que LR et LRO ont une précision très faible, car ce sont des modèles naïfs sans interactions entre caractéristiques. CL, TAN et LRT obtiennent des résultats très satisfaisants. LRT traite les interactions entre caractéristiques pour chaque classe, tandis que TAN les traite à l'aide d'une structure d'arbre commune estimée à partir des deux distributions p et q. Le modèle CL est très similaire au modèle LRT en termes de structure et d'interactions apprises, mais ces deux modèles diffèrent sur l'apprentissage des paramètres.

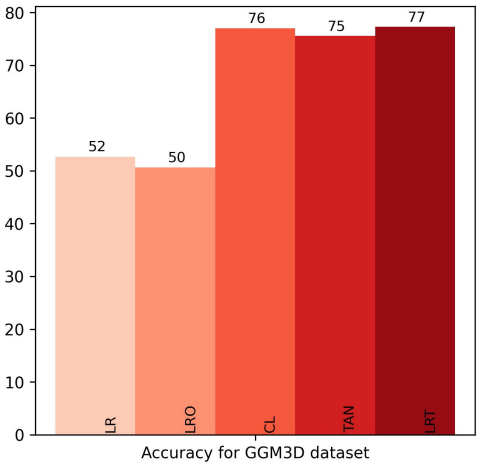


Figure 1: Histogram of accuracy on the test set for GGM3D.

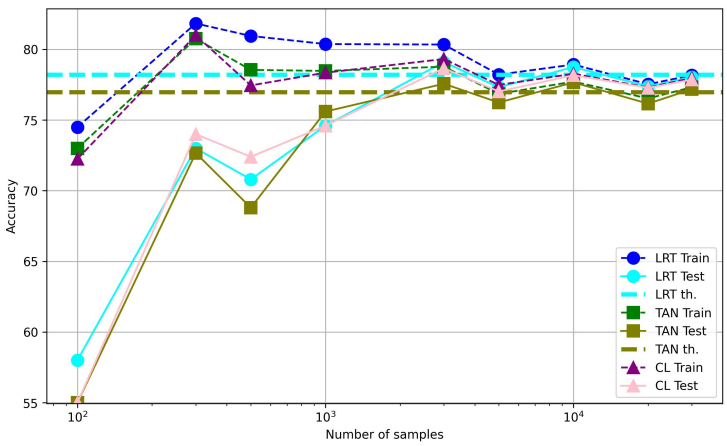


Figure 2: Comparaison of accuracy between train and test sets for GGM3D.

La deuxième expérience se concentre sur la convergence des méthodes LRT et TAN. Les deux méthodes se comportent, une fois de plus, de manière très similaire. Sur la Fig. 2, nous observons d'abord un régime de surapprentissage pour un faible nombre d'échantillons, suivi d'une convergence légèrement inférieure aux performances des classifieurs théoriques. Cela s'explique simplement par la discrétisation.

La dernière expérience vise à distinguer TAN et LRT (Tableau 1). TAN construit un modèle d'arbre global en priorisant les corrélations les plus fortes, sans spécificité de classe. Pour mettre en évidence cette faiblesse, nous simulons notre ensemble de données GGM20D à l'aide de matrices de corrélation en arbre Σ_p et Σ_q, où les valeurs de corrélation par paire sont identiques pour toutes les arêtes. Étant donné que toutes les corrélations sont également importantes mais que les structures des matrices diffèrent, TAN perd nécessairement des informations précieuses dans son choix des arêtes. LRT, utilise à la fois les interactions entre caractéristiques communes aux deux classes et celles spécifiques à chaque classe, ce qui lui permet d'accumuler davantage d'informations à mesure que la corrélation globale augmente.

Corrélation	0.05	0.1	0.3	0.5	0.6	0.9
TAN	53.898	57.728	68.664	74.2	79.592	78.698
LRT	55.166	60.566	78.822	89.804	92.874	96.838

Table 1: Précision des modèles en fonction de la corrélation (LRT en rouge, TAN en orange)

5 Conclusion

Dans cet article, nous montrons qu'un encodage approprié des variables explicatives permet d'entraîner une régression logistique qui correspond au test Bayésian optimal pour choisir entre deux distributions modélisées chacune par un réseau Bayésien sous forme d'arbre. Nous proposons également un encodage dit suffisant qui permet d'estimer de façon unique les paramètres du modèle.

References

[1] Friedman et al., "Bayesian network classifiers," *Machine learning*, vol. 29, pp. 131–163, 1997.

[2] Chickering et al., "Large-sample learning of bayesian networks is np-hard," *Journal of Machine Learning Research*, vol. 5, pp. 1287–1330, 2004.

[3] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[4] Amin Zollanvari, "Nonoptimality of the maximum-weight dependence tree in classification," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 71–75, 2016.

[5] Baptiste Schall, Rodolphe Anty, and Lionel Fillatre, "One-hot logistic regression for radiomics-based classification," in *2024 IEEE International Conference on Image Processing (ICIP 2024)*, 2024.

[6] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag New York, 2nd edition, 1994.

Apprentissage par renforcement et plans d'expérience

Apprentissage actif basé sur le gradient des processus gaussiens pour l'analyse de sensibilité globale.

Guerlain Lambert*[1], Céline Helbert[1], and Claire Lauvernet[2]

[1]École Centrale de Lyon (ECL) – Université de Lyon – 36 avenue Guy de Collongue - 69134 Ecully cedex, France

[2]INRAE RiverLy – Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France

Résumé

Ces dernières décennies, l'expérimentation numérique s'est imposée comme une alternative aux essais sur le terrain, souvent coûteux, pour étudier les phénomènes physiques et évaluer l'impact des activités humaines sur l'environnement. Toutefois, la simulation numérique peut s'avérer complexe en raison de la modélisation détaillée des phénomènes, impliquant des temps de calcul prolongés et une paramétrisation soignée. Pour remédier à ces difficultés, le recours à un méta-modèle construit à partir d'observations du modèle complet permet de réduire considérablement la charge computationnelle. Dans ce contexte, nous proposons une approche d'apprentissage actif qui, pour un budget d'évaluations identique, cible intelligemment les zones informatives du modèle afin d'optimiser les performances de l'apprentissage. Plus précisément, notre démarche s'appuie sur les avancées récentes en apprentissage actif pour l'analyse de sensibilité (indices de Sobol, DGSM), qui exploitent les dérivées issues d'un processus gaussien (GP). En étudiant la distribution jointe a posteriori du gradient du GP, nous développons une méthodologie plus complète et robuste que l'approche existante. Les performances de cette approche seront évaluées sur plusieurs cas d'étude, allant de modèles synthétiques à une application environnementale réelle portant sur la diffusion de pesticides.

Mots-Clés: Apprentissage actif, Analyse de sensibilité, Processus gaussien, Métamodélisation.

*Intervenant

Intervalle de confiance asymptotique pour une probabilité estimée à partir d'un plan hypercube latin

Théo Sylvestre [1,5], Guillaume Damblin [2], Amandine Marrel [3,5], Raksmy Nop [4]

[1] *Université Paris-Saclay, CEA, Service de Génie Logiciel pour la Simulation, 91190, Gif-sur-Yvette, France, theo.sylvestre@cea.fr*
[2] *Université Paris-Saclay, CEA, Service de Génie Logiciel pour la Simulation, 91190, Gif-sur-Yvette, France, guillaume.damblin@cea.fr*
[3] *CEA, DES, IRESNE, DER, Cadarache, France, amandine.marrel@cea.fr*
[4] *Université Paris-Saclay, CEA, Service de Thermo-hydraulique et de Mécanique des Fluides, 91191, Gif-sur-Yvette, France, raksmy.nop@cea.fr*
[5] *LMA Université d'Avignon, EA 2151, 84029 Avignon, France*

Résumé : En simulation numérique, l'estimation de la probabilité qu'une réponse dépasse un certain seuil nécessite un plan d'expériences bien choisi afin de limiter le coût des calculs. Dans le cadre des approches non adaptatives, l'échantillonnage par hypercube latin est couramment utilisé car il permet notamment une réduction de la variance asymptotique par rapport aux plans Monte Carlo traditionnels. Cependant, dans le contexte d'estimation d'une probabilité, les méthodes actuelles ne permettent pas d'estimer la variance asymptotique de l'estimateur associé à ce type d'échantillonnage, en raison de la nature discontinue de la fonction indicatrice 0-1. Pour remédier à cette limite, nous proposons pour estimer cette variance une approche novatrice basée sur les indices de Sobol issus de l'analyse de sensibilité. Compte tenu du contexte "given-sample", les indices de Sobol sont estimés à l'aide des estimateurs de Chatterjee appliqués aux données disponibles. Enfin, nous illustrons notre approche sur un exemple numérique.

Mots-clés : Plan hypercube latin, Estimateur d'une probabilité, Variance asymptotique, Indices de Sobol, Estimateur de Chatterjee

Abstract : In computer experiments, estimating the probability that a response exceeds a certain threshold requires a well-chosen experimental design to limit computational costs. In the context of non-adaptive approaches, a commonly used technique is the Latin hypercube sampling (LHS), which enables asymptotic variance reduction compared to traditional Monte Carlo designs. However, in this context of probability estimation, current methods do not allow for the estimation of the asymptotic variance of the LHS-based estimator, due to the discontinuous nature of the 0-1 indicator function. To overcome this limitation, we propose to estimate this variance a novel approach based on Sobol' indices from sensitivity analysis. Within the given-sample context, the estimation of Sobol' indices relies on Chatterjee's estimators. Then, we illustrate our approach through a numerical example.

Keywords. Latin hypercube sampling, Probability estimator, Asymptotic variance, Sobol' indices, Chatterjee estimator

Introduction

Dans le cadre de l'étude des scénarios accidentels sur les réacteurs nucléaires, on s'intéresse à l'estimation par simulation d'une probabilité d'occurrence d'un évènement critique. Cette estimation repose sur le choix d'un plan d'expériences pour sélectionner les réalisations des variables d'entrée en lesquelles appeler le simulateur. Ce choix influence directement les biais et variances des estimateurs considérés. Dans le cadre de l'échantillonnage non-adaptatif, un plan d'expériences très répandu est l'échantillonnage par hypercube latin (*Latin hypercube sampling*, LHS), introduit par McKay et al. (1979). À budget de simulation fixé, ce type de plan offre une meilleure couverture de l'espace des variables d'entrée par rapport aux plans aléatoires simples (*Simple random sampling*, SRS), permettant une potentielle réduction de la variances des estimateurs statistiques. En particulier, à taille de plan fixée, McKay et al. (1979) propose une expression explicite de la relation entre les variances des plans LHS et SRS. De son coté, Stein (1987) établit un théorème central limite dans le cadre du plan LHS, fournissant une expression de la variance asymptotique associée. Cependant, dans le cas d'un estimateur de probabilité basé sur un unique plan LHS, il est difficile d'utiliser la formule introduite par McKay et al. (1979), qui repose sur des probabilités conditionnelles. Également, les méthodes proposées dans la littérature ne permettent pas l'estimation de la variance asymptotique proposée par Stein (1987), en raison de la nature discontinue des quantités impliquées. Pour réussir à estimer cette variance, nous proposons ici une nouvelle approche basée sur les indices d'analyse de sensibilité de Sobol, estimés par les estimateurs de Chatterjee. Nous commencerons par exposer la théorie permettant l'estimation de la variance asymptotique d'un estimateur de probabilité venant d'un LHS. Nous illustrerons ensuite la procédure d'estimation sur un cas jouet permettant de juger de la performance de notre méthode.

1 Estimateur empirique d'une probabilité

Soit Y un code de simulation prenant en entrée $K \in \mathbb{N}^\star$ variables d'entrée indépendantes notées $\mathbf{X} = (X_1, \cdots, X_K) \in \mathbb{R}^K$, et α un seuil donné. Nous cherchons à estimer la probabilité p de dépassement de ce seuil α, qui s'écrit :

$$p = \mathbf{E}\left(\mathbf{1}_{Y(\mathbf{X})>\alpha}\right).$$

Considérons $n \in \mathbb{N}$ ainsi qu'un n-échantillon i.i.d $(\mathbf{X}^{(1)}, ..., \mathbf{X}^{(n)})$ de la variable aléatoire $\mathbf{X}$. Un estimateur Monte-Carlo de la probabilité cible p est donné par la moyenne empirique :

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y(\mathbf{X}^{(i)})>\alpha} \tag{1}$$

Cet estimateur dépend des réalisations échantillonnées des variables $(\mathbf{X}^{(i)})_{i\in[\![1,n]\!]}$. Des alternatives existent à l'échantillonnage aléatoire simple (SRS), visant notamment à contrôler la manière dont est réalisé l'échantillonnage de sorte à améliorer sa représentativité et réduire l'erreur d'estimation. Parmi ces méthodes, on s'intéresse ici à l'échantillonnage par LHS.

2 Plans hypercubes latins

Notons $n \in \mathbb{R}$ la taille du plan considéré. Notons Σ_n le groupe de permutations de l'ensemble des entiers $\{1, 2, \ldots, n\}$. Une matrice $D_{\mathbf{X}} \in M_{n,K}(\mathbb{R})$ est un plan hypercube latin si chacune de ses colonnes est construite comme une permutation particulière $\sigma \in \Sigma_n$, où K est la dimension du problème.

Supposons que pour $1 \leq j \leq K$, chaque marginale X_j suit indépendemment une distribution uniforme sur $[0, 1]$. Soit $\left\{\sigma_1^X, \sigma_2^X, \cdots, \sigma_K^X\right\}$ un ensemble de K permutations indépendantes de Σ_n. Pour $1 \leq i \leq n$, générons les échantillons uniformes :

$$U_j^{(i)} \underset{i.i.d.}{\sim} \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right).$$

Ensuite, le plan

$$D_{\mathbf{X}} = \left[\frac{\sigma_1^X - 0.5 + U_1^{\sigma_1^X}}{n}, \cdots, \frac{\sigma_i^X - 0.5 + U_j^{\sigma_j^X}}{n}, \cdots, \frac{\sigma_K^X - 0.5 + U_K^{\sigma_K^X}}{n}\right] \in \mathcal{M}_{n,K}([0, 1])$$

est appelé un *plan hypercube latin aléatoire*, tel que pour $1 \leq j \leq K$:

$$\frac{\sigma_j^X - 0.5 + U_j^{\sigma_j^X}}{n} = \left(\frac{\sigma_j^X(1) - 0.5 + U_j^{(\sigma_j^X(1))}}{n}, \cdots, \frac{\sigma_j^X(n) - 0.5 + U_j^{(\sigma_j^X(n))}}{n}\right)^T \in \mathbb{R}^n.$$

Graphiquement, le LHS revient à découper chaque marginale en n intervalles équiprobables et à générer des points dans certains hyperrectangles formés par ces subdivisions, de manière à obtenir l'uniformité des projections sur chaque dimension. Si X_j ne suit pas la loi uniforme sur $[0, 1]$, il est possible d'utiliser la méthode dite "par inversion de la fonction de répartition" pour satisfaire la loi de tirage.

Comme pour un plan SRS, le LHS conduit à des estimateurs sans biais (i.e. $\mathbf{E}(\hat{p}) = p$). Mais la variance associée à l'estimateur (1) est différente. En effet, on peut relier la variance des estimateurs du plan LHS avec ceux du plan SRS. D'après McKay et al. (1979), on a :

$$\mathrm{Var}(\hat{p}_{LHS}) = \mathrm{Var}(\hat{p}_{SRS}) + \left(\frac{n-1}{n}\right) \frac{1}{n^K (n-1)^K} \sum_R (\mu_i - \mu)(\mu_j - \mu) \tag{2}$$

avec R l'ensemble des cellules du plan LHS n'ayant pas de coordonnées en commun et $\mu_i := \mathbb{P}(Y(\mathbf{X}) > \alpha | \mathbf{X} \in cell_i)$. La notation $cell_i$ permet de noter chaque cellule issue du partitionnement de l'espace par le plan LHS. Un avantage du plan LHS est qu'il permet de réduire la variance des estimateurs sous certaines conditions. Par exemple, quand la fonction étudiée $\mathbf{X} \mapsto \mathbf{1}_{Y(\mathbf{X})>\alpha}$ est monotone, ou quand le nombre d'observations est très largement supérieur à la dimension du problème, le second terme de l'équation (2) est négatif, comme montré dans McKay et al. (1979). Cela signifie dans ce cas que l'estimateur basé sur un plan LHS a une variance plus faible que l'estimateur basé sur un plan SRS. Un gain de précision des estimations est ainsi obtenu sans faire appel à des méthodes de construction séquentielle de plans d'expériences (i.e. échantillonnage adaptatif). Or, en pratique, un plan LHS ne fournit au maximum qu'un point par cellule $cell_i$, ce qui rend l'estimation des probabilités conditionnelles μ_i difficile et complique donc l'utilisation de l'équation (2). Il est donc pertinent de développer une approche alternative permettant d'estimer la variance de l'estimateur (1).

3 Variance asymptotique et théorème central limite

D'après Stein (1987), en utilisant le reste non principal R de la décomposition fonctionnelle de Hoeffding pour $h : \mathbf{X} \mapsto \mathbf{1}_{Y(\mathbf{X})>\alpha}$, il est possible d'exprimer la variance asymptotique de l'estimateur (1) issu d'un plan LHS. Précisément, puisque h est de carré intégrable, R peut être exprimé pour $\mathbf{X} = (X_1, ..., X_K) \in \mathbb{R}^K$ comme :

$$R(\mathbf{X}) = h(\mathbf{X}) - \mathbf{E}(h(\mathbf{X})) - \sum_{j=1}^{K} \big(\mathbf{E}(h(\mathbf{X})|X_j) - \mathbf{E}(h(\mathbf{X}))\big) \tag{3}$$

La variance asymptotique de l'estimateur est alors donnée par :

$$\mathrm{Var}(\hat{p}_{LHS}) = \frac{1}{n} \int R^2(\mathbf{X}) dF + \underset{n\to+\infty}{o}\left(\frac{1}{n}\right) \tag{4}$$

où F représente la mesure de probabilité de $\mathbf{X}$.

Par ailleurs, Stein (1987) propose également un théorème central limite (TCL) qui s'exprime comme suit :

$$\sqrt{n}(\hat{p}_{LHS} - p) \underset{n\to+\infty}{\longrightarrow} \mathcal{N}\left(0, \int R^2(\mathbf{X}) dF\right) \tag{5}$$

Ce TCL nous permet de déduire des intervalles de confiance (IC) asymptotiques. De plus, comme $\int \big(\mathbf{E}(h(\mathbf{X})|X_j) - \mathbf{E}(h(\mathbf{X}))dF_j = 0$ (F_j étant la mesure de probabilité de X_j) et $\int R(\mathbf{X}) dF_{-j} = 0$ (F_{-j} étant la mesure jointe de $\mathbf{X}$ privé de la j-ième composante), on obtient $\mathbf{E}(R(\mathbf{X})) = 0$ et $\int R^2(\mathbf{X}) dF = \mathrm{Var}(R(\mathbf{X}))$.

Par ailleurs, comme les variables X_j et X_k sont supposées indépendantes pour $j \neq k$, il en découle que les variables aléatoires $\mathbf{E}(h(\mathbf{X})|X_j)$ et $\mathbf{E}(h(\mathbf{X})|X_k)$ sont aussi indépendantes. Finalement, on obtient :

$$\mathrm{Var}(h(\mathbf{X})) = \sum_{j=1}^{K} \mathrm{Var}(\mathbf{E}(h(\mathbf{X})|X_j)) + \mathrm{Var}(R(\mathbf{X})). \tag{6}$$

Ainsi, d'après l'équation (6), si on parvient à estimer les termes $\big(\mathrm{Var}(\mathbf{E}(h(\mathbf{X})|X_i)\big)_{j\in[\![1,K]\!]}$, on pourra calculer $\mathrm{Var}(R(\mathbf{X}))$, la variance asymptotique de l'estimateur (1) issu d'un plan LHS, ainsi que son IC à partir de l'équation (5). Si les variables aléatoires $(\mathbf{E}(h(\mathbf{X})|X_j))_j$ étaient continues, nous aurions pu utiliser des méthodes de régression comme le propose Owen (1992) pour les estimer et en déduire leur variance. Mais leur discontinuité héritée de celle de h dans le cas de l'estimation d'une probabilité ne permet pas d'utiliser cette approche. Une alternative est alors d'estimer directement les variances conditionnelles $\mathrm{Var}(\mathbf{E}(h(\mathbf{X})|X_j))$, qui à un facteur près, correspondent aux indices de sensibilité de Sobol.

4 Indices de Sobol du premier ordre

En se plaçant dans le même formalisme que précédemment, les indices de Sobol du premier ordre sont définis par :

$$S^1_{X_j} = \frac{\text{Var}(\mathbf{E}(h(\mathbf{X})|X_j)}{\text{Var}(h(\mathbf{X}))} \quad \text{pour } j \in \{1, ..., K\} \tag{7}$$

Ces indices quantifient l'influence des entrées $(X_j)_{j\in[\![1,K]\!]}$ sur $h(\mathbf{X})$ et il est possible de d'exprimer la variance de R (Eq. (3)) à partir de ces indices :

$$\int R^2(\mathbf{X})dF = \text{Var}(R(\mathbf{X})) = \left(1 - \sum_{j=1}^{K} S^1_{X_j}\right) \text{Var}(h(\mathbf{X})) \tag{8}$$

Finalement, en remarquant que $\text{Var}(h(\mathbf{X})) = p(1-p)$, on peut construire un estimateur de la variance asymptotique LHS de $\hat{p}$ à partir des estimateurs de $S^1_{X_j}$ (notés $\widehat{S^1_{X_j}}$) et de $\hat{p}$.

$$\widehat{\text{Var}(R(\mathbf{X}))} = \left(1 - \sum_{j=1}^{K} \widehat{S^1_{X_j}}\right) \hat{p}(1-\hat{p}) \tag{9}$$

Il reste maintenant à estimer $S^1_{X_j}$ dans un contexte "given-sample". Pour cela, on propose d'utiliser les estimateurs de Chatterjee introduits par Gamboa et al. (2023).

5 Estimateur de Chatterjee

Supposons que l'on dispose d'un n-échantillon i.i.d $\left(\left(\mathbf{X}^{(1)}, h\left(\mathbf{X}^{(1)}\right)\right), \ldots, \left(\mathbf{X}^{(n)}, h\left(\mathbf{X}^{(n)}\right)\right)\right)$ de $(\mathbf{X}, h(\mathbf{X}))$, avec $n \in \mathbb{N}$. Pour chaque $j \in [\![1, K]\!]$, on considère le couple de variables aléatoires $(X_j, h(\mathbf{X}))$ et on note $\pi_j : i \mapsto \pi_j(i)$ la fonction représentant le rang de $X_j^{(i)}$ dans le n-échantillon. Pour l'estimateur étudié, si k est tel que $\pi_j(k) = n$, alors π doit vérifier $\pi_j(k)+1 = 1$. On suppose aussi que la loi de X_j est diffuse [1]. L'estimateur de Chatterjee de $S^1_{X_j}$ pour le n-échantillon est alors défini par :

$$\widehat{S^{1,C}_{X_j}} = \frac{\frac{1}{n}\sum_{i=1}^{n} h\left(\mathbf{X}^{(\pi_j(i))}\right) h\left(\mathbf{X}^{(\pi_j(i)+1)}\right) - \left(\frac{1}{n}\sum_{i=1}^{n} h\left(\mathbf{X}^{(i)}\right)\right)^2}{\frac{1}{n}\sum_{i=1}^{n} h\left(\mathbf{X}^{(i)}\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n} h\left(\mathbf{X}^{(i)}\right)\right)^2} \tag{10}$$

À partir d'un unique échantillon SRS, il est ainsi possible d'estimer les indices de Sobol du premier ordre. Cet estimateur est généralement faiblement biaisé et il est aussi montré dans Gamboa et al. (2023) qu'il est consistant pour un plan SRS, i.e. $\widehat{S^{1,C}_{X_j}} \underset{n\to\infty}{\to} S^1_{X_j}$. Dans le cas d'un

1. Pour une loi diffuse, la probabilité d'obtenir une valeur particulière est nulle. En pratique, il est donc supposé impossible d'avoir deux réalisations identiques dans l'échantillon considéré.

LHS, la consistance a été observée sur plusieurs exemples numériques ; il serait intéressant de la démontrer théoriquement en étendant les travaux de Gamboa et al. (2023). Pour en déduire la consistance de l'estimateur $\widehat{\mathrm{Var}(R(\mathbf{X}))}$ (défini avec $\widehat{S^{1,C}_{X_j}}$), il sera nécessaire de démontrer en plus la nullité asymptotique des termes de covariance entre les estimateurs.

6 Illustration de la procédure d'estimation sur un cas jouet

Afin de tester numériquement les performances de notre estimateur, on considère le cas test analytique suivant :

$$Y : \mathbf{X} \mapsto 1_{X_1 \geq s_1} 1_{X_2 \geq s_2} \tag{11}$$

avec $\mathbf{X} = (X_1, X_2) \sim \mathcal{U}[0, 100]^2$, $X_1 \perp X_2$ et $(s_1, s_2) \in \mathbb{R}^2$. La zone de défaillance paramétrée par (s_1, s_2) est définie par $Y(\mathbf{X}) = 1$.

Pour illustrer les performances de notre estimateur de la variance asymptotique, on considère le cas $s_1 = s_2 = 80$ et on construit un intervalle de confiance (IC) à 95% pour la probabilité p. Pour différentes tailles n, on calcule les IC obtenus à partir d'un LHS (et de notre estimateur de la variance) et les taux de couverture observés de ces IC (idéalement égal à 95%). Ces derniers sont comparés à ceux d'un plan SRS. La Figure 1 présente les résultats obtenus. En complément, la Figure 2 illustre la largeur moyenne des IC à 95% pour les deux types de plan.

On remarque sur la Figure 1 que les taux de couverture des deux plans sont similaires et très proches de la cible fixée à 95%. Cela tend à valider la méthode proposée pour l'estimation de la variance asymptotique. De plus, la Figure 2 montre que les largeurs des IC par plan LHS sont plus faibles que celles par plan SRS. L'approche par LHS combinée à notre estimateur de la variance asymptotique permet ainsi d'obtenir des IC asymptotiques plus précis, i.e. plus étroits tout en étant du bon niveau.

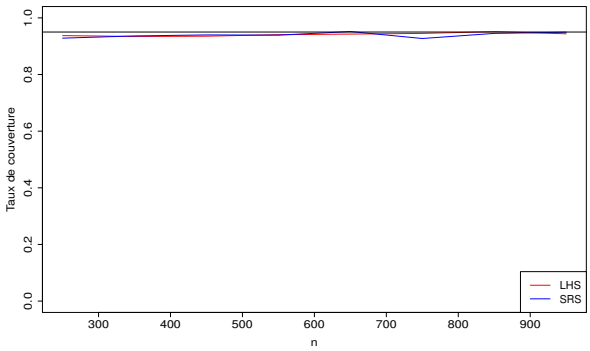


FIGURE 1 – Taux de couverture des IC à 95% en fonction de la taille du plan n. La ligne noire horizontale correspond à la cible de couverture de 95%.

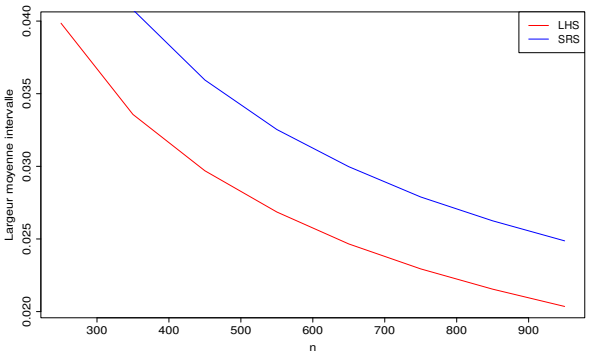


FIGURE 2 – Largeur moyenne des IC à 95% en fonction de la taille n du plan.

Cet exemple encourage l'application de la méthode à des simulations réelles. En particulier,

cette méthodologie a été appliquée au code de thermohydraulique CATHARE [2] dans le cadre de la sûreté nucléaire. Le code prend en entrée 9 variables physiques ($K = 9$) et renvoie le flux thermique dans le coeur du réacteur. Grâce à la méthodologie présentée, nous avons pu estimer la probabilité de dépassement d'un seuil critique identifié pour la sûreté ainsi que l'IC asymptotique associé.

7 Conclusion

La méthodologie présentée, spécialement adaptée aux plans LHS, permet une estimation asymptotique de la variance d'un estimateur de probabilité. Innovante, cette approche repose sur l'utilisation des indices de Sobol et des estimateurs de Chatterjee. Elle élargit les possibilités d'utilisation des plans LHS, permettant de bénéficier de leur réduction de variance asymptotique dans des problématiques d'analyse de risque. Elle a été appliquée sur un simulateur thermohydraulique pour estimer la probabilité que le flux critique dépasse un seuil critique. En perspective, il serait intéressant de démontrer théoriquement la consistance de l'estimateur de la variance asymptotique qui est proposé, en étendant au cadre LHS les travaux de Gamboa et al. (2023) sur la consistance des estimateurs de Chatterjee.

Références

F. Gamboa, P. Gremaud, T. Klein, and A. Lagnoux. 2023. Global Sensitivity Analysis : a new generation of mighty estimators based on rank statistics. arXiv :2003.01772 [math.ST] https://arxiv.org/abs/2003.01772

M.D. McKay, R.J. Beckman, and W.J. Conover. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 42, 1 (1979), 55–61. https://www.jstor.org/stable/1268522

A.B. Owen. 1992. A Central Limit Theorem for Latin Hypercube Sampling. *Journal of the royal statistical society series b-methodological* 54 (1992), 541–551. https://api.semanticscholar.org/CorpusID:121890318

M. Stein. 1987. Large sample properties of simulations using Latin hypercube sampling. *Technometrics* 29, 2 (1987), 143–151. https://www.jstor.org/stable/1269769

2. https://cathare.cea.fr/

Planification d'expériences séquentielle en inférence bayésienne objective, application à l'estimation de courbes de fragilité sismique

Antoine Van Biesbroeck [1]

[1] *CMAP École polytechnique, Institut Polytechnique de Paris & Université Paris-Saclay, CEA, France,* antoine.van-biesbroeck@polytechnique.edu

Résumé. Nous proposons une stratégie de planification d'expériences séquentielle dans un contexte d'inférence bayésienne. Notre stratégie prend la forme d'un apprentissage incrémentiel maximisant séquentiellement une quantité d'information, mesurée par un indice de sensibilité. Le caractère non subjectif du canevas bayésien est également mis en avant, avec la considération d'un prior de référence "proprement contraint". Ce dernier, défini dans des travaux récents, constitue un prior qui bénéficie du caractère objectif des priors de référence, tout en assurant un aspect propre du posterior. Cette propreté est essentielle pour la mise en œuvre de la planification d'expérience. Notre méthode est évaluée dans le contexte de l'estimation de courbes de fragilité sismique. Sur un cas d'étude pratique provenant de l'industrie nucléaire, nos résultats démontrent la capacité de la méthode à rapidement atteindre le biais irréductible du modèle choisi.

Mots-clés. Inférence bayésienne, prior de référence, planification d'expériences

Abstract. We propose a strategy for sequential design of experiments in a Bayesian inference context. Our strategy takes the form of online learning that sequentially maximizes a quantity of information, which is measured by a sensitivity index. The non-subjective nature of the Bayesian framework is also emphasized, with the consideration of a "properly constrained" reference prior. The latter, defined in recent work, constitutes a prior that benefits from the objective character of reference priors, while ensuring a proper aspect of the posterior. This proper aspect is essential for the implementation of experimental design. Our method is evaluated in the context of estimating seismic fragility curves. Evaluated on a practical case study taken from the nuclear industry, our results demonstrate the method's ability to quickly reach the irreducible bias of the chosen model.

Keywords. Bayesian inference, reference prior, design of experiments

1 Introduction

Le canevas bayésien est souvent plébiscité dans des problèmes pratiques d'inférence statistique où il apporte de la régularité tout en permettant une quantification des incertitudes. Cependant, il reste tributaire du choix de sa loi a priori, qui centralise beaucoup de questions et de critiques : (i) un prior trop informatif contrevient à la confiance en les estimations, (ii) un prior peu informatif donne souvent lieu à des posteriors intractables et une incertitude immense. Tels sont les "trous" du canevas bayésien selon Gelman & Yao (2020).

Pour traiter le problème (i), la théorie des priors de référence propose la définition de priors dits “objectifs”, permetttant une auditabilité des estimations (Berger et al., 2009). Ces priors sont généralement non-informatifs et impropres (ils ne s’intègrent pas). Qui plus est, dans de nombreux cas d’études leur “faible informativité” donne lieu à des posteriors impropres, ce qui place l’étude au cœur du problème (ii).

Dans ce travail, nous proposons une méthode d’inférence bayésienne “objective” qui réussit à éviter le problème (ii). Notre méthode s’adresse à des modèles statistiques de prédictions, elle se base sur deux éléments essentiels.

D’abord, nous présentons des résultats théoriques récents (Van Biesbroeck, 2024) qui étendent la théorie des priors de référence pour définir un prior de référence “proprement contraint”, qui est objectif et donne lieu à un posterior propre. Ce prior s’appuie sur une contrainte paramétrée et il reste en pratique essentiel de s’assurer que cette contrainte n’influe pas sur les estimations. Ensuite, nous proposons une méthodologie de planification d’expériences séquentielle, qui vient maximiser l’information apportée au posterior par les observations. Cette démarche assure l’informativité du posterior (permettant d’éviter le problème (ii)) ainsi que le caractère objectif de cette information (permettant d’éviter le problème (i)), puisqu’elle provient des observations.

A la fin de cette communication, nous présentons l’application de cette méthode ainsi que ses performances dans un contexte d’estimation de courbes de fragilité sismique.

2 Contexte d’inférence bayésienne

2.1 Vraisemblance d’un modèle expérience-résultat

Ce travaille se concentre sur le cas d’un système ‘entrée-sortie’ : modélisé comme la réalisation d’un couple de variables aléatoires $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. On définit un tel système par une observation en deux temps : on observe

1. l’entrée : une réalisation de la variable X, elle suit une loi $\mathbb{P}_X$ connue,
2. la sortie : une réalisation de la variable Y, selon sa loi conditionnelle à X : $\mathbb{P}_{Y|X}$.

On suppose une modélisation paramétrique de la loi conditionnelle $\mathbb{P}_{Y|X}$ écrite sous le paradigme bayésien : il existe une variable aléatoire $\theta \in \Theta \sim \pi$ telle que $\mathbb{P}_{Y|X,\theta}$ est connue et admet une densité $p(y|x, \theta)$.

La loi π s’appelle le prior, après observation de k réalisations indépendantes de (X, Y) notées $(\mathbf{x}^k, \mathbf{y}^k)$, on définit la vraisemblance ℓ_k et le posterior $p(\theta|\mathbf{x}^k, \mathbf{y}^k)$ selon

$$\ell_k(\mathbf{y}^k|\mathbf{x}^k, \theta) = \prod_{i=1}^{k} p(y_i|x_i, \theta) \; ; \quad p(\theta|\mathbf{x}^k, \mathbf{y}^k) = \frac{\pi(\theta)\ell_k(\mathbf{y}^k|\mathbf{x}^k, \theta)p_{\mathbf{X}^k}(\mathbf{x}^k)}{p_{\mathbf{Y}^k}(\mathbf{y}^k)}, \tag{1}$$

où $p_{\mathbf{X}^k}(\mathbf{x}^k)$ (resp. $p_{\mathbf{Y}^k}(\mathbf{y}^k)$) désigne la loi de k répétitions i.i.d. de X (resp. de Y). Il est à noter que l’explicitation à une constante près de la loi a posteriori $p(\theta|\mathbf{x}^k, \mathbf{y}^k)$ ne dépend pas des lois de X et de Y, mais seulement du prior et de la vraisemblance.

2.2 Prior de référence proprement contraint

La définition du prior reste une étape essentielle pour la construction d'estimateurs de θ a posteriori. Il s'agit d'une question ouverte pour laquelle nous prenons le parti de nous appuyer sur la théorie des priors de référence (Berger et al., 2009). Celle-ci propose un cadre pour définir un prior comme étant "objectif". Le prior de référence π^* est alors définit comme une solution du problème d'optimisation $\pi^* \in \arg\max_{\pi\in\mathcal{P}} \mathscr{I}^k(\pi)$ lorsque $k \longrightarrow \infty$, où $\mathscr{I}^k$ est l'information mutuelle et $\mathcal{P}$ est une classe de prior (typiquement l'ensemble des priors à densité continue). Le problème d'optimisation asymptotique est défini comme suit :

$$\pi^* \in \underset{\pi\in\mathcal{P};\, k\to\infty}{\arg\max}\, \mathscr{I}^k(\pi) \iff \forall \pi \in \mathcal{P},\ \lim_{k\to\infty} \mathscr{I}^k(\pi^*) - \mathscr{I}^k(\pi) \geq 0. \tag{2}$$

L'information mutuelle $\mathscr{I}^k$ est quant à elle écrite $\mathscr{I}^k(\pi) = \mathbb{E}_{\mathbf{X}^k,\mathbf{Y}^k}[D_\delta(\pi(\cdot)||p(\cdot|\mathbf{X}^k,\mathbf{Y}^k)]$ où D_δ désigne une δ-divergence : $D_\delta(p||q) = \int f_\delta\left(\frac{p(x)}{q(x)}\right) q(x)dx$ avec $f_\delta(x) = \frac{x^\delta - 1}{\delta(\delta-1)}$, $\delta \in (0,1)$. Le prior de référence est dans beaucoup de cas le prior de Jeffreys (Clarke & Barron, 1994), que nous notons J. Pour rappel, $J(\theta) \propto \sqrt{|\det \mathcal{F}(\theta)|}$, $\mathcal{F}$ désignant l'information de Fisher.

Le défaut du prior de Jeffreys est qu'il est souvent impropre. Si la vraisemblance est mal spécifiée, il peut donner lieu à un posterior toujours impropre, contrevenant à toute estimation a posteriori de θ. Pour contrer cette problématique, nous proposons de nous appuyer sur des travaux récents (Van Biesbroeck, 2024), qui définissent un prior de référence "proprement contraint" en proposant l'ajout d'une contrainte au problème d'optimisation (2) :

$$\tilde{\pi}^* \in \underset{\substack{\pi\in\mathcal{P};\, k\to\infty \\ \text{sujet à } \mathbb{E}_\pi g<\infty}}{\arg\max}\ \mathscr{I}^k(\pi), \text{ où } g \text{ est telle que } \int_\Theta J\cdot g^{1/\delta} < \infty \text{ et } \int_\Theta J\cdot g^{1+1/\delta} < \infty. \tag{3}$$

La solution du problème ci-dessus est un prior propre, définit par $\tilde{\pi}^* \propto J\cdot g^{1/\delta}$. Le choix de g doit être pensé comme une correction des taux asymptotiques trop lourds de J. Cependant, il est essentiel de la choisir de sorte à ce que la contrainte "$\mathbb{E}_\pi g < \infty$" restreigne peu l'espace $\mathcal{P}$, afin de ne pas dénaturer l'aspect objectif du prior.

3 Méthodologie de planification d'expériences séquentielle

Nous définissons dans cette section une stratégie de planification d'expériences pour sélectionner les variables d'entrée plutôt que de les tirer aléatoirement selon $\mathbb{P}_X$. Il faut tout de même noter que la connaissance de $\mathbb{P}_X$ reste indispensable, puisqu'elle intervient dans l'expression du prior de référence explicité en section précédente. Dans ce travail nous suggérons de maximiser une quantité d'information : celle apportée par les données à la loi a posteriori. Cette information est mesurée par le moyen d'un indice de sensibilité : en supposant avoir observé $(\mathbf{x}^k, \mathbf{y}^k)$, l'entrée suivante observée x_{k+1} est sélectionnée selon

$$x_{k+1} = \underset{x\in\mathcal{X}}{\arg\max}\, \mathcal{I}_{k+1}(x), \quad \text{où} \quad \mathcal{I}_{k+1}(x_{k+1}) = \mathbb{E}_{y_{k+1}|\mathbf{x}^k,\mathbf{y}^k}[D_\delta(p(\cdot|\mathbf{x}^k,\mathbf{y}^k)||p(\cdot|\mathbf{x}^{k+1},\mathbf{y}^{k+1}))].$$

Nous remarquons que l'expression de l'indice $\mathcal{I}$ ressemble à celle de l'information mutuelle $\mathscr{I}$, les deux s'appuyant sur une démarche de maximisation d'information apportée par les observations. L'indice $\mathcal{I}$ est approximé par estimateurs de Monte-Carlo. Pour la bonne définition

de $\mathcal{I}$ il est essentiel que les probabilités a posteriori soient bien définies et propres. D'où l'intérêt d'avoir à disposition un prior proprement construit.

4 Application à l'estimation de courbes de fragilité sismique

Les courbes de fragilité sismique représentent un outil clé d'aide à la décision en génie parasismique (Kennedy et al., 1980). Elles définissent la probabilité qu'un équipement mécanique donné tombe en panne conditionnellement à une mesure d'intensité d'un signal sismique. Cette mesure d'intensité, abrégée IM, est un scalaire calculé à partir du signal sismique. Nous décrivons dans la suite un modèle d'estimation de ces courbes de fragilité.

4.1 Modèle probit-lognormal

Le problème d'estimation de courbes de fragilité se modélise comme un problème d'entrée-sortie tel que décrit en section 2.1 : $X \in \mathcal{X} = (0, \infty)$ désigne ici l'IM du signal sismique, et $Y \in \mathcal{Y} = \{0, 1\}$ vaut 1 si l'équipement est en panne à l'issue du séisme, ou 0 sinon. On considère la modélisation paramétrique suivante de la loi conditionnelle $\mathbb{P}_{Y|X}$:

$$p(y|x, \theta) = P_f^\theta(x)^y (1 - P_f^\theta(x))^{1-y}, \text{ où } \theta = (\alpha, \beta) \in (0, \infty)^2, \; P_f^\theta(x) = \Phi(\beta^{-1} \log(x/\alpha)), \quad (4)$$

Φ désignant la fonction de répartition d'une gaussienne centrée réduite. La loi de X est généralement assimilée à une loi log-normale dans ce contexte : $\log X \sim \mathcal{N}(\mu_X, \sigma_X^2)$. Dans ce cas, Van Biesbroeck et al. (2024) ont étudié les comportements asymptotiques du prior de Jeffreys pour ce modèle.

La vraisemblance du modèle (4) donne souvent lieu à des posteriors impropres, surtout avec peu de données. Il est donc essentiel de construire un prior propre, comme suggéré en section 2.2. Aussi, on veillera à ce que la méthode de planification d'expériences donne lieu une loi des observations qui limite ce phénomène d'impropreté.

La contrainte dans l'équation (3) écrite avec $g(\theta) = \beta^\varepsilon$, $\varepsilon \in (0, \frac{2\delta}{1+\delta})$ est peu restrictive, et satisfait la condition demandée. La solution, s'écrit alors $\tilde{\pi}^*_\varepsilon \propto J(\theta)\beta^{\varepsilon/\delta}$. Elle peut être paramétrée selon $\gamma = \varepsilon/\delta$, donnant lieu à un éventail de priors candidats :

$$\tilde{\pi}^*_\gamma(\theta) \propto J(\theta)\beta^\gamma, \quad \gamma \in (1, 2/(1+\delta)) \subset (0, 2).$$

Tant que $0 < \gamma < 2$, ce prior est propre et donne lieu à un posterior propre. Le cas limite $\gamma = 0$ revient au prior de Jeffreys, qui donne lieu à un posterior parfois impropre (en fonction des données observées). Le cas extrême $\gamma = 2$ ne donne jamais lieu à un posterior propre.

4.2 Résultats sur un cas d'étude issu de l'industrie nucléaire

Présentation du cas d'étude Le système mécanique étudié est le même que celui décrit dans (Van Biesbroeck et al., 2024). Il s'agit du circuit de tuyauterie représenté dans la figure

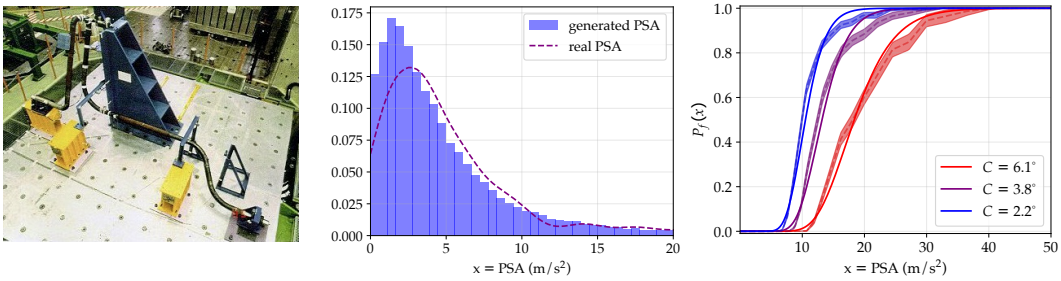


FIGURE 1 – A gauche : maquette de la tuyauterie étudiée. Au centre : distribution des PSA calculés à partir des signaux sismiques réels (en noir) et artificiels (en bleu). A droite : courbes de fragilité de référence pour différents seuils de défaillance.

1-gauche. Des simulations numériques sur une base de signaux sismiques artificiels donnent lieu pour ce système à une base de donnée conséquente ($M = 8 \cdot 10^4$) de résultats de son comportement sous séisme. Les données que nous considérons prennent la forme de paires (IM, Réponse). L'IM du séisme pris en compte est l'accélération pseudo-spectrale (PSA) à 5Hz, conformément aux caractéristiques de l'équipement. C'est elle qui définit l'entrée X. La réponse Y vaut 1 si la structure est en panne à l'issue du séisme. La panne arrive lors d'une rotation excessive du premier coude de la tuyauterie au delà d'un certain seuil C.

Une courbe de fragilité de référence P_f^{ref} est définie par une estimation non-paramétrique (Trevlopoulos et al., 2019), à partir des M données. On définit aussi la courbe P_f^{MLE} comme la courbe de fragilité issue du modèle probit (4) avec θ estimé par maximimisation de la vraisemblance sur les M données. Ces deux courbes sont affichées en figure 1-droite pour différent seuils de rotation excessive de la tuyauterie. Dans la suite nous considérons seulement le seuil $C = 3.8°$. La différence $\|P_f^{\text{ref}} - P_f^{\text{MLE}}\|_{L^2}^2$ est appelée biais carré de modèle irréductible.

Résultats numériques Dans la figure 2, nous comparons les résultats de deux méthodes : la méthode standard (tirage des séismes selon leur loi originale, et estimation bayésienne avec le prior de Jeffreys), ainsi que la méthode de planification d'expériences (P.E.) pour chaque $\gamma \in \{0, 0.1, 0.3, \ldots, 1.9\}$. Le cas $\gamma = 0$ est particulier car la P.E. ne peut pas être faite dans ce cas. Il correspond alors à la conduite de la P.E. avec $\gamma = 0.1$ pour la sélection des entrées, suivie du remplacement de $\tilde{\pi}_{0.1}^*$ par $\tilde{\pi}_0^*$ pour les estimations a posteriori.

Pour chacune des méthodes, nous avons réalisé 100 réplications de la même expérience : générer un échantillon d'observations de taille $k_{\max} = 250$, et calculer pour chaque $k = 10, 20, \ldots, k_{\max}$ la loi a posteriori $p(\theta|\mathbf{x}^k, \mathbf{y}^k)$. Il s'en suit le cacul des valeurs moyennes de l'erreur quadratique $\mathcal{E}$ (figure 2-gauche), du bais carré $\mathcal{B}$ (figure 2-centre) et de la largeur de crédibilité $\mathcal{W}$ (figure 2-droite) en fonction de k. Ces trois métriques sont définies par :

$$\mathcal{E}^{|\mathbf{x}^k,\mathbf{y}^k} = \mathbb{E}_{\theta|\mathbf{x}^k,\mathbf{y}^k}[\|P_f^\theta - P_f^{\text{ref}}\|_{L^2}^2], \quad \mathcal{B}^{|\mathbf{x}^k,\mathbf{y}^k} = \|m^\theta - P_f^{\text{ref}}\|_{L^2}^2, \quad \mathcal{W}^{|\mathbf{x}^k,\mathbf{y}^k} = \|q_{1-r/2}^\theta - q_{r/2}^\theta\|_{L^2}^2,$$

où la norme carrée $\|P\|_{L^2}^2 = \int_{x_1}^{x_2} P(x)^2 dx$ est évaluée par la méthode de Simpson ; P_f^θ désigne un processus de $\mathcal{X}$ dans $[0, 1]$ dont la loi est la mesure image de la loi de θ par l'application $x \mapsto \Phi(\beta^{-1} \log \frac{x}{\alpha})$; pour tout x, la médiane de $P_f^\theta(x)$ est notée $m^\theta(x)$ et son r-quantile est noté $q_r^\theta(x)$. Ici, on prend $r = 0.05$.

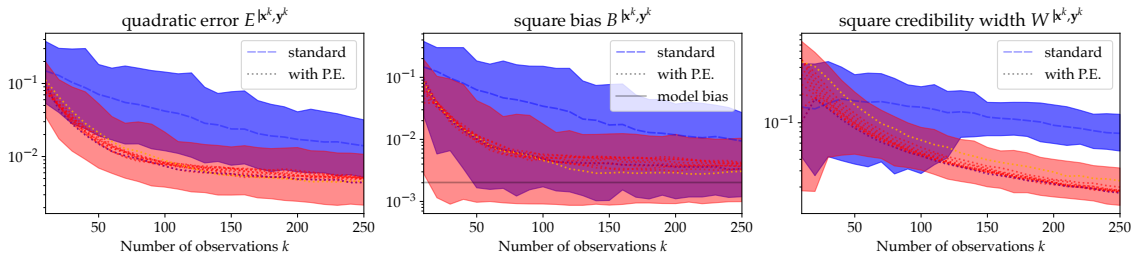


FIGURE 2 – En traits bleus : erreurs moyennes dans le cas standard entourées de leurs intervalles de confiance à 95%. En pointillés : les erreurs moyennes pour les différentes méthodes P.E. entourées par le plus large de leurs intervalles de confiance à 95%. Deux tracés sont particuliers : les cas $\gamma = 0$ (en violet) et $\gamma = 1.9$ (en orange).

Commentaire des résultats et conclusion Les résultats de la figure 2 témoignent à la fois de la robustesse et de la performance de la méthode. Tout d'abord, en comparant les résultats provenant des différentes méthodes de P.E., on remarque que les performances sont très similaires quel que soit γ : l'objectivité du prior n'est pas remise en question par ce paramètre. Ensuite, la sur-performance de la P.E. est principalement évidente lorsque le nombre de données observées est faible. On remarque que rapidement, l'estimation issue de la P.E. atteint le bais de modèle irréductible. Les observations supplémentaires donnent alors lieu à une réduction de l'intervalle de crédibilité autour d'une médiane biaisée.

Références

Berger, J. O., Bernardo, J. M., & Sun, D. (2009). The formal definition of reference priors. *Ann. Stat.*, 37(2).

Clarke, B. S. & Barron, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Stat. Plan. Inference*, 41(1).

Gelman, A. & Yao, Y. (2020). Holes in Bayesian statistics. *JPhysG*, 48(1).

Kennedy, R. P., Cornell, C. A., Campbell, R. D., Kaplan, S. J., & Perla, H. F. (1980). Probabilistic seismic safety study of an existing nuclear power plant. *Nucl. Eng. Des*, 59(2).

Trevlopoulos, K., Feau, C., & Zentner, I. (2019). Parametric models averaging for optimized non-parametric fragility curve estimation based on intensity measure data clustering. *Struct. Saf.*, 81.

Van Biesbroeck, A. (2024). Properly constrained reference priors decay rates for efficient and robust posterior inference. *arXiv.2409.13041*.

Van Biesbroeck, A., Gauchy, C., Feau, C., & Garnier, J. (2024). Reference prior for Bayesian estimation of seismic fragility curves. *Probab. Eng. Mech.*, 76.

Une approche policy-gradient pour des actions ordinales

Simon Weinberger [1,2] & Jairo Cugliari [2]

[1] *Essilor*
[2] *Laboratoire ERIC, Université Lumière Lyon 2, ERIC,* jairo.cugliari@univ-lyon2.fr

Résumé. En apprentissage par renforcement, la paramétrisation softmax est l'approche standard pour les politiques sur des espaces d'actions discrètes. Cependant, elle ne prend pas en compte la relation d'ordre entre les actions. Motivés par un problème industriel réel, nous proposons une nouvelle paramétrisation des politiques basée sur les modèles de régression ordinale adaptés au cadre de l'apprentissage par renforcement. Notre approche répond aux défis pratiques, et des expériences numériques démontrent son efficacité dans des applications réelles et dans des tâches à actions continues, où la discrétisation de l'espace d'actions combinée à la politique ordinale offre des performances compétitives.

Mots-clés. Apprentissage par renforcement, Policy-gradient, Régression ordinale.

Abstract. In reinforcement learning, the softmax parametrization is the standard approach for policies over discrete action spaces. However, it fails to capture the order relationship between actions. Motivated by a real-world industrial problem, we propose a novel policy parametrization based on ordinal regression models adapted to the reinforcement learning setting. Our approach addresses practical challenges, and numerical experiments demonstrate its effectiveness in real applications and in continuous action tasks, where discretizing the action space and applying the ordinal policy yields competitive performance.

Keywords. reinforcement Learning, Policy-gradient, Ordinal Regresion.

1 Introduction

L'apprentissage par renforcement (RL) permet de résoudre des problèmes nécessitant des décisions séquentielles afin de maximiser une récompense à long terme, suivant une politique à déterminer. Des algorithmes tels que *Trust Region Policy Optimization* (TRPO, Schulman *et al.*, 2015), *Proximal Policy Optimization* (PPO, Schulman *et al.*, 2017) ou *Natural Policy Gradient* (NPG, Kakade, 2001) estiment des politiques paramétrées qui induisent une distribution sur l'espace d'actions. Toutefois, la majorité des applications reposent sur une distribution logistique (*softmax*) pour les actions catégoriques (Agarwal *et al.*, 2020) ou une distribution normale pour les actions continues (Sutton et Barto, 2018), négligeant la relation d'ordre entre les actions ordinales.

Cette limitation affecte des problèmes réels où les actions suivent un ordre naturel. Par exemple, les verres électrochromiques d'ESSILOR, qui proposent quatre niveaux de teinte

ordonnés ajustés en fonction de la lumière ambiante via un capteur (*Ambient Light Sensor*, ALS) et des seuils prédéfinis. Nous proposons un algorithme de contrôle de teinte pour des lunettes électrochromiques, ajustant la teinte parmi des classes ordonnées allant d'une classe claire (i.e. verres transparents) à la clase plus obscure (des lunettes de soleil foncées).

Nous proposons une politique ordinale inspirée des modèles de régression ordinale linéaire (Agresti, 2010) et des réseaux de neurones CORAL (Cao *et al.*, 2020). Cette approche permet l'utilisation de méthodes d'optimisation par gradient de politique telles que REINFORCE, NPG, TRPO et PPO. Nos résultats expérimentaux montrent que, dans des contextes où les actions sont ordonnées, les politiques ordinales convergent plus rapidement et vers de meilleures solutions que les politiques multinomiales classiques (i.e. *softmax*). De plus, elles offrent des performances comparables à celles des politiques continues lorsque l'espace d'actions continues est discrétisé de manière appropriée.

Le processus d'apprentissage consiste en la collecte des données suivantes : les capteurs mesurent l'état, une teinte est proposée (action), et l'utilisateur l'accepte ou en choisit une autre (récompense). L'objectif est d'optimiser la politique pour maximiser les récompenses actualisées, tout en tenant compte des éventuels délais de réaction. Ce processus est itéré pendant un épisode (par exemple, une journée), permettant des mises à jour dynamiques du contrôle de la teinte.

2 Cadre théorique

Nous considérons un processus de décision markovien où $\mathcal{S}$, $\mathcal{A}$ et $\mathcal{R} = \mathbb{R}$ représentent les espaces d'états, d'actions et de récompenses, respectivement. L'évolution du système est régie par une densité (inconnue) de transition d'état :

$$p(s', r|s, a) := \mathbb{P}(S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a), \qquad s, s' \in \mathcal{S}, a \in \mathcal{A}, r \in \mathcal{R}.$$

À chaque moment (ou étape) t, les actions sont prises selon une politique, $\pi(.|s_t)$, que nous considérons étant une distribution de probabilité paramétrique, c'est-à-dire nous posons $\pi = \pi_\theta$ où le vecteur de paramètres $\theta \in \Theta \subset \mathbb{R}^d$. Pour un facteur d'actualisation donné $\gamma \in [0, 1[$, une action $a \in \mathcal{A}$, un état $s \in \mathcal{S}$ et une politique π, nous définissons la fonction de valeur d'état-action $Q^\pi(s, a)$ et la fonction de valeur d'état $V^\pi(s)$ comme étant respectivement :

$$Q^\pi(s, a) = \mathbb{E}_{\pi_\theta}\left(\sum_{t=0}^{\infty} \gamma^t R_t | S_0 = s, A_0 = a\right) \quad \text{et} \quad V^\pi(s) = \mathbb{E}_{\pi_\theta}\left(\sum_{t=0}^{\infty} \gamma^t R_t | S_0 = s\right).$$

L'objectif de l'optimisation de politique est de trouver $\theta \in \Theta$ qui maximise la fonction objective J, définie comme :

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left(\sum_{t=0}^{\infty} \gamma^t R_t\right).$$

3 Politiques ordinales

Nous considérons l'espace des actions $\mathcal{A}$ un ensemble fini d'éléments $\{a_k : k = 1, \ldots, K\}$, où K est un nombre naturel supérieur ou égal à 3, et il existe une relation d'ordre entre les actions : $a_1 < a_2 < \ldots < a_{K-1} < a_K$. Sans perte de généralité, nous identifions les éléments avec leurs étiquettes $\mathcal{A} = [\![1, K]\!]$.

Paramétrisation Soit une fonction $g_\omega \colon \mathcal{S} \to \mathbb{R}$ paramétrée par un vecteur de poids $\omega \in \Omega$, un ensemble de $K-1$ seuils ordonnés $\tau_1 < \tau_2 < \ldots < \tau_{K-2} < \tau_{K-1}$ et la fonction sigmoïde $\sigma(x) = (1+\exp(-x))^{-1}$. Nous vodrions une politique qui choisira l'action $A = j$ si la quantité latente $A^* = g_\omega(s) + e, e \sim \text{Logistique}(0, 1)$, se trouve entre les seuils $\tau_{j-1} < A^* \leq \tau_j$. Cette formulation classique en régression ordinale est équivalente à l'expression

$$\mathbb{P}(A \leq j | S = s) = \sigma\left(\tau_j - g_\omega(s)\right). \tag{1}$$

En utilisant la convention $\tau_0 = -\infty$ et $\tau_K = \infty$, l'équation (1) induit une distribution de probabilité sur l'espace des actions $\mathcal{A}$ et qu'on appellera *politique ordinale*, définie par

$$\pi(a|s) = \sigma(\tau_a - g_\omega(s)) - \sigma(\tau_{a-1} - g_\omega(s)), a \in \mathcal{A}.$$

Intuitivement, la fonction g_ω est une forme de score, plus elle est élevée plus la politique choisira actions élevées. L'espace paramétrique de la politique ordinale est $\Theta = \Omega \times \Delta_{K-1}$ où Δ_{K-1} est l'ensemble de seuils ordonnées.

4 Expériences numériques

Données simulées La mesure ALS est simulée par un processus gaussien. La réponse utilisateur est modélisée par une politique ordinale inconnue π_U. À un instant donné, une classe a_t est proposée pour l'état s_t, et un score de désagrément Z_t est mis à jour selon :

$$Z_{t+1} = (1 - \pi_U(a_t|s_t))^{\gamma_r} + \gamma_d Z_t.$$

L'utilisateur réagit avec une probabilité $\sigma(Z_{t+1})$ et, le cas échéant, sélectionne une classe selon $\pi_U(.|s_t)$. Le paramètre $\gamma_r \in \mathbb{R}^+$ contrôle la fréquence des réactions : une valeur faible entraîne des ajustements fréquents, tandis qu'une valeur élevée signifie que l'utilisateur n'intervient que lorsque la classe proposée est improbable sous π_U. Le paramètre $\gamma_d \in [0, 1]$ modélise la mémoire de l'utilisateur, une valeur nulle indiquant une réaction uniquement basée sur la classe actuelle.

Apprentissage L'espace des états $\mathcal{S}$ correspond aux mesures du ALS, l'espace des actions $\mathcal{A}$ aux quatre classes de teinte, et l'espace des récompenses $\mathcal{R} = \{-3, -2, -1, 0\}$ à la différence absolue entre la teinte proposée et celle choisie par le porteur. Nous comparons les politiques ordinales et softmax mises à jour avec REINFORCE, NPG et TRPO. Nous utilisons $\gamma_r = 0.5$,

$\gamma_d = 1$ et un taux d'actualisation de 0.9. Chaque épisode comprend soixante étapes et se conclut par une mise à jour de la politique. Pour chaque combinaison de stratégie de mise à jour et de politique, nous simulons quatre cents épisodes, générant ainsi dix trajectoires d'apprentissage par configuration.

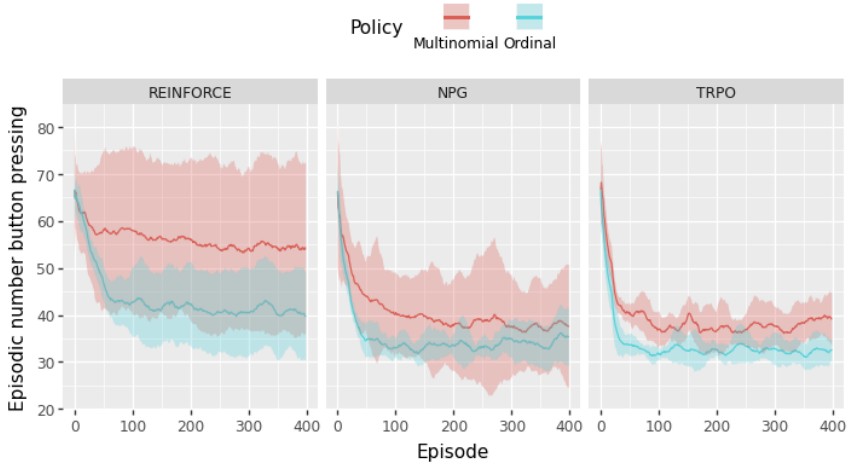


Figure 1: Récompense totale par épisode en fonction du nombre d'épisodes, pour les méthodes REINFORCE, NPG et TRPO et les politiques ordinale et multinomial. Les courbes moyennes sont calculées sur dix tirages aléatoires, avec un lissage par moyenne glissante sur une fenêtre de vingt épisodes. Les zones ombrées représentent la moyenne ± un écart-type.

Résultats Les politiques ordinales convergent plus rapidement et vers de meilleures solutions que les politiques multinomiales. De plus, cette paramétrisation se révèle plus stable, comme en témoigne la diminution de l'écart-type lors de l'utilisation d'une politique ordinale (voir Figure 1). Parmi les six méthodes testées, la mise à jour d'une politique ordinale avec TRPO produit les meilleurs résultats, offrant une amélioration plus rapide que REINFORCE et une stabilité légèrement supérieure à celle de NPG.

Discrétisation d'un espace d'actions continues Les actions ordinales peuvent émerger de la discrétisation d'un espace d'actions continues. Considérons un espace borné $\mathcal{A} = [m, M]$, que nous discrétisons selon : $\mathcal{A}_K = \{a_k\}_{k=1}^K$ avec $a_k = m + k\frac{M-m}{K}$. L'ensemble $\mathcal{A}_K$ est naturellement ordonné et possède un nombre fini d'actions.

Dans le cas d'actions continues multivariées, la discrétisation peut être appliquée séparément à chaque dimension. Nous implémentons cette approche en utilisant une politique ordinale avec 17 classes par dimension pour résoudre divers environnements de référence en RL, notamment `Mujoco` (Todorov *et al.*, 2012). L'optimisation de la politique est réalisée via PPO, avec l'implémentation des actions continues de Huang *et al.*, 2022, et les mêmes hyperparamètres.

Nous utilisons le réseau de neurones implémenté pour l'algorithme PPO en actions continues dans Huang *et al.*, 2022, basé sur un MLP à deux couches pour paramétrer $g_\omega(\cdot)$, ainsi qu'un écart-type à estimer par dimension d'action, indépendant de l'état. Les expériences

sont menées dans les environnements du Tableau 1, avec un taux d'apprentissage de $3 \cdot 10^{-4}$ et un facteur d'actualisation $\gamma = 0.99$ pour PPO en actions continues et PPO ordinal. Les résultats montrent que la politique ordinale atteint des performances équivalentes, voire légèrement supérieures, à celles de la politique continue dans tous les environnements testés.

Environnement	$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{A}\rvert$	*Environnement*	$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{A}\rvert$
`Ant-v4`	105	8	`BipedalWalker-v3`	24	4
`HalfCheetah-v4`	17	6	`Hopper-v4`	11	3
`Humanoid-v4`	348	17	`InvertedDoublePendulum-v4`	9	1
`Pusher-v4`	23	7	`Walker2d-v4`	17	6

Table 1: Environnements à actions continues où des actions ordinales ont été utilisées via la discrétisation.

5 Conclusion

L'essor des dispositifs portables hautement configurables ouvre la voie à des modèles adaptatifs répondant aux besoins individuels. Le RL fournit un cadre puissant pour ce type de problèmes, mais son déploiement en conditions réelles nécessite de prendre en compte les contraintes spécifiques à chaque application. Lorsqu'un système contrôle un paramètre ordonné, l'adoption d'une politique ordinale garantit une robustesse structurelle : quelles que soient les mises à jour, la politique conserve la hiérarchie des actions, assurant ainsi une meilleure cohérence.

Enfin, la discrétisation des actions continues demeure un enjeu ouvert. Dans la Section 4, nous avons fixé arbitrairement le nombre de classes et la méthode de discrétisation, mais une approche plus adaptative pourrait être explorée. Par exemple, au fil de l'apprentissage, de nouvelles classes pourraient être ajoutées dynamiquement en conservant le prédicteur appris g_ω et en introduisant de nouveaux seuils, optimisant ainsi la granularité de la politique.

Bibliographie

Agarwal, A., Kakade, S. M., Lee, J. D., et Mahajan, G. (2020) Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, PMLR, pp. 64-66.

Huang, S., Dossa, R.F.J., Ye, C., Braga, J., Chakraborty, D., Mehta, K., Araújo, J.G.: Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. Journal of Machine Learning Research 23(274), 1–18 (2022),

Kakade, S. M. (2001). A natural policy gradient. Advances in neural information processing systems, 14.

Schulman, J., Levine, S., Abbeel, P., Jordan, M. et Moritz, P. (2015) Trust Region Policy Optimization. *Proceedings of the 32nd International Conference on Machine Learning*,

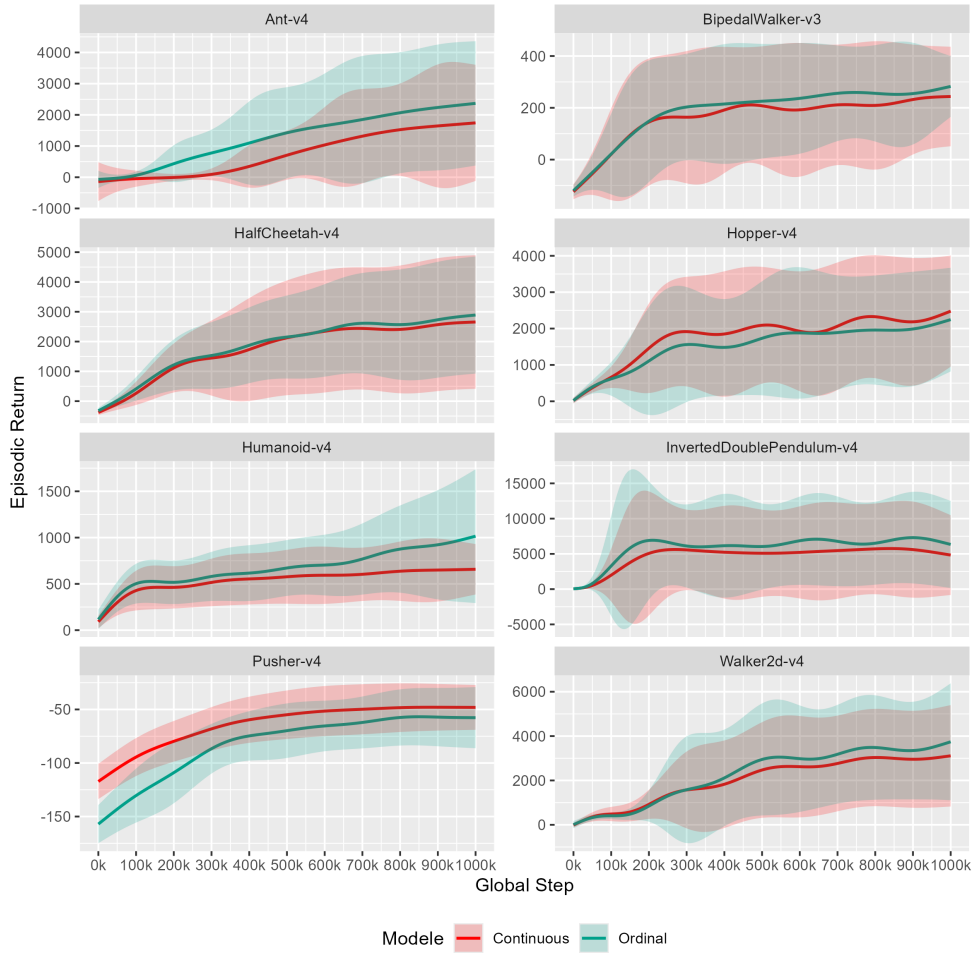


Figure 2: Courbes d'apprentissage pour différentes politiques sur divers environnements. Moyenne et intervalle de confiance à 95%, ajustés à l'aide d'un modèle GAM gaussien à échelle localisée sur des données générées avec cinq tirages aléatoires par environnement.

PMLR

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., et Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint `arXiv:1707.06347`.

Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. A Bradford Book, Cambridge, MA, USA (2018)

Todorov, E., Erez, T., Tassa, Y.: Mujoco: A physics engine for model-based control. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 5026–5033. IEEE (2012).

Apprentissage sur données déséquilibrées

Échantillonnage en deux étapes pour la classification déséquilibrée à l'aide de forêts aléatoires centrées infinies

Moria Mayala *[1], Erwan Scornet [1], Charles Tillier [2] & Olivier Wintenberger [1]

[1] *Laboratoire de Probabilités, Statistiques et Modélisations – Sorbonne Université*
[2] *Laboratoire de Mathématiques – Université de Versailles Saint-Quentin-en-Yvelines*

Résumé. Les données déséquilibrées dans les taches de classification sont aujourd'hui identifiées comme un problème majeure en apprentissage automatique. De nombreuses solutions ont été proposées pour relever les défis posés par les données déséquilibrées, allant des approches au niveau des données aux approches au niveau des algorithmes. Nous étudions la classification avec sous-échantillonnage, qui est l'une des approches populaires au niveau des données utilisant les forêts aléatoires centrées infinies(ICRF). Nous établissons notamment un théorème central limite pour cet estimateur ICRF sous certaines hypothèses de régularité sur la fonction de régression. Nous montrons que si cette méthode résout efficacement le déséquilibre, elle introduit un biais substantiel par rapport à l'approche classique des forêts aléatoires. Pour résoudre ce problème, nous proposons une proceduire d'échantillonnage en deux étapes afin de reduire le biais asymptotiques de ICRF. Une courte étude de simulation illustre les performances de la méthode proposée.

Mots-clés. Classification binaire, données déséquilibrées, forêts aléatoires centrées infinies.

Abstract. Imbalanced data in classification tasks has been identified as a top problem in machine learning. Numerous solutions have been proposed to address the challenges posed by imbalanced data, ranging from data-level approaches to algorithm-level approaches. We study classification with undersampling, which is one of the most popular data-level approaches using Infinite Centered Random Forests (ICRFs). We specifically establish a central limit theorem for this ICRF estimator under certain regularity assumptions on the regression function. We show that while this method effectively addresses the imbalance, it introduces a substantial bias compared to the classical random forest approach. To address this issue, we propose a two-step sampling procedure to reduce the asymptotic bias of ICRFs. A short simulation study illustrates the performance of the proposed method.

Keywords. binary classification, class imbalance, infinite centered random forests.

1 Introduction

Il est connu que la prédiction sur les données déséquilibrées répresente une difficulté majeure en apprentissage automatique, en particulier dans les modèles de classification, qu'il

s'agisse de classification binaire (ex :détecter une maladie, détection d'images, intrusion réseau, détection de fraude) ou de classification multi-classes (ex : prédire le modèle de voiture acheté), pour plus de détailles une bonne revue a été proposée par Jason Brownlee (2020). Ainsi, proposer une procedure éfficace d'estimation dans ce cadre est un enjeu d'intérêt. Nous présentons ici le cas binaire, qui est plus simple à appréhender et peut ensuite facilement se généraliser au multi-classes.

Le problème de base est le suivant. On considère $(X_i, Y_i)_{1\leq i\leq n}$ des observations i.i.d de même loi que le couple $(X, Y) \in [0,1]^d \times \{0,1\}$. La loi jointe de (X, Y) est complètement déterminer par la loi marginale de X et la fonction de régression μ est définie par

$$\mu(\mathbf{x}) := \mathbb{P}(Y = 1|X = \mathbf{x}), \qquad \mathbf{x} \in [0,1]^d. \tag{1}$$

Notre objectif est d'estimer μ lorsque les classes $\{0,1\}$ ne sont pas représentées de manière égale. Pour quantifier le degré de *déséquilibre*, nous utilisons la notion de ratio de déséquilibre (Imbalance ratio) définie comme suit

$$IR = n_0/n_1$$

où n_0 et n_1 désignent respectivement le cardinal des échantillons majoritaire et minoritaire, respectivement. Il ya un déséquilibre dans les données lorsque $IR > 1$. Dès lors que la valeur $IR \approx 1$ correspondant à des données parfaitement équilibrées (cf. Robert O'Brien et al (2019)). Sans perte de généralité, nous supposons que la classe minoritaire correspond à l'etiquette 1.

2 Méthode

Introduit par Leo Breiman (2001), les forêts aléatoires constituent un algorithme d'apprentissage très populaire qui offre d' excellente performances, et une grande flexibilité dans sa capacité à gérer tous les types de données. Ainsi dans le contexte des données déséquilibrées (imbalanced data), Leo Breiman et al (2004) proposent une nouvelle variante de forêts aléatoires, simplifiées appelée forêts aléatoires équilibrées qui traite à tous points de vue le probleme d'imbalanced data en sous-échantillonnant la classe majoritaire afin d'améliorer les performances de classification par rapport à la classe minoritaire. Sans perte de généralité, les forêts aléatoires centrées étudiées par Biau (2012) interviennent dans le même contexte en vue de relever les défis associés à la prédiction d'événements rares. Il est important ici de rappeler la procédure de construction des arbres aléatoires centrés: Tous les nœuds de l'arbre sont associés à des cellules rectangulaires de sorte qu'à chaque étape de la construction de l'arbre, l'ensemble des cellules associées aux feuilles de l'arbre (c'est-à-dire, les nœuds externes) forme une partition de $[0,1]^d$. La racine de l'arbre est $[0,1]^d$ elle-même. La procédure suivante est ensuite répétée $\lceil \log_2 \omega_s \rceil$ fois, où $\log_2$ est le logarithme en base 2, $\lceil \cdot \rceil$ la fonction plafond et $\omega_s \geq 2$ un paramètre déterministe, fixé au préalable par l'utilisateur, et dépendant éventuellement de n.

1. À chaque nœud, une coordonnée de $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$ est sélectionnée, la j-ème caractéristique ayant une probabilité $1/d$ d'être sélectionnée.

2. À chaque nœud, une fois la coordonnée sélectionnée, la division se fait au milieu du côté choisi.

Notre approche s'appuit sur les travaux antérieurs de Wager et Athey (2018) et de Peng et al. (2022) qui ont établi la normalité asymptotique de variantes infinies similaires de différents estimateurs à l'aide des outils de U-statistique. Notre objectif est de construire un estimateur $\widehat{\mu}_s^{\mathrm{ICRF}}$ de μ. Etant donné, $Z_i := (X_i, Y_i)$, pour $i \in I := \{1, \ldots, n\}$ des couples d'observation i.i.d de même loi que le couple (X, Y) comme susmentionné. Á partir du schéma de construction des arbres centrés, nous sommes donc prêts à construire l'estimateur ICRF. Considérons $S \subset I$ comme un ensemble d'indices de taille $|S| = s$ avec $1 \leq s \leq n$ et notons $\mathbf{Z}_S := (Z_{i_1}, \ldots, Z_{i_s})$ le vecteur de taille s à valeur dans $([0,1]^d \times \{0,1\})$ construit uniformément sans remplacement à partir de s individus dans $\mathbf{Z}_I$. Un arbre individuel$T^s(\mathbf{x}; \mathbb{U}; \mathbf{Z}_S)$ évalué en un point $\mathbf{x} \in [0,1]^d$ basé sur le sous-échantillon $\mathbf{Z}_S$ s'écrit

$$T^s(\mathbf{x}; \mathbb{U}; \mathbf{Z}_S) := \frac{\sum_{i \in S} Y_i \mathbb{1}\{X_i \in L_{\mathbb{U}}(\mathbf{x})\}}{N_{L_{\mathbb{U}}(\mathbf{x})}(\mathbf{X}_S)} \tag{2}$$

où

$$N_{L_{\mathbb{U}}(\mathbf{x})}(\mathbf{X}_S) := |\{i \in S : X_i \in L_{\mathbb{U}}(\mathbf{x})\}| = \sum_{i \in S} \mathbb{1}\{X_i \in L_{\mathbb{U}}(\mathbf{x})\}$$

est le nombre d'éléments de l'échantillon partiel S qui appartiennent à la feuille $L_{\mathbb{U}}(\mathbf{x})$. Par convention, nous fixons $\frac{0}{0} = 0$ signifiant que la prédiction de l'arbre $T^s(\mathbf{x}; \mathbb{U}; \mathbf{Z}_S)$ défini dans Equation (2) est arbitrairement égale à zéro lorsqu'il n'y a pas d'observation qui tombe dans la feuille $L_{\mathbb{U}}(\mathbf{x})$. L'estimateur ICRF noté $\widehat{\mu}_s^{\mathrm{ICRF}}$ est alors défini comme la somme contant $\binom{n}{s}$ éléments sur tous les sous-ensembles $S \subset I$ avec cardinal $|S| = s$ par

$$\widehat{\mu}_s^{\mathrm{ICRF}}(\mathbf{x}) := \binom{n}{s}^{-1} \sum_{S \subset I, |S| = s} \mathbb{E}[T^s(\mathbf{x}; \mathbb{U}; \mathbf{Z}_S) \mid \mathbf{Z}_S].$$

Pour établir nos résultats asymptotiques, nous nous appuyons sur les hypothèses suivantes concernant la distribution des covariables et la régularité de la fonction de régression.

(H0) La covariable X est uniformément distribuée sur $[0,1]^d$.

(H1) La fonction de régression μ est une fonction L-lipschitzienne.

Théorème 1 *Supposons que les hypothèses **(H0)** et **(H1)** soient vérifiées. Si de plus* $n/(s(\log \omega_s)^{(d-1)/2}) \to \infty$ *et* $\omega_s / \max(n^{d/(d+1)}, s^2/n) \to \infty$ *quand* $n \to \infty$*, il existe une constante* $C(d) > 0$ *telle que pour* $\mathbf{x} \in [0,1]^d$*, nous avons*

$$\sqrt{\frac{n(\log(\omega_s))^{(d-1)/2}}{\omega_s}} (\widehat{\mu}_s^{\mathrm{ICRF}}(\mathbf{x}) - \mu(\mathbf{x})) \xrightarrow{d} \mathcal{N}(0, C(d)\mu(\mathbf{x})(1 - \mu(\mathbf{x}))), \qquad n \to \infty.$$

3 Résultats numériques

Dans cette section, nous donnons à l'aide de données synthétiques, une illustration numérique de nos résultats théoriques. Nous comparons notre approche avec la méthode classique de forêts aléatoires. Pour mener ces expériences, nous utilisons les paramètres par défaut fournis par le package `randomForest` de `R` (voir Liaw (2002)).

Distributions de Génération de Données Synthétiques. Nous commençons par générer un ensemble d'entraînement $z = \{(x_i, y_i)\}_{i=1}^{n}$ à partir de

$$\begin{cases} X \sim U([0,1]^d)\,, \\ Y|X = x \sim \mu(x)\,, \end{cases} \tag{SI}$$

avec $d \geq 2$. Les étiquettes Y suivent une distribution de Bernoulli, et la fonction de régression est définie comme

$$\text{Modèle 1} : \mu(x) := 1/(1 + \exp(-h(x_1, x_2))),$$

où $h(x_1, x_2) = 3x_1 + 2x_2$. Nous considérons le scénario de grand déséquilibre dans l'ensemble de données, à savoir 90% de la classe majoritaire ($Y = 0$) contre 10% de la classe minoritaire ($Y = 0$). Le processus d'évaluation est mené en utilisant 30 répétitions de Monte Carlo, pour plusieurs échantillons de taille $n = \{100, \ldots, 600\}$.

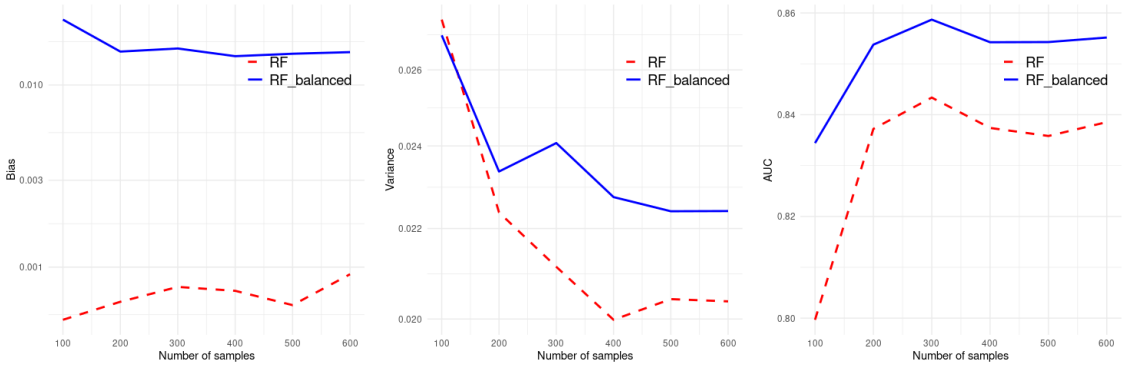


Figure 1: Comparaison de la tendance de l'AUC avec celle de la décomposition du biais-variance en fonction de la taille des échantillons.

4 Conclusion

Dans cette étude, nous avons examiné l'estimateur ICRF et démontré un théorème central limite pour cet estimateur sous certaines hypothèses de régularité concernant la fonction de régression. Une analyse plus approfondie de sa variance et des intervalles de confiance bootstrap pourrait offrir des garanties supplémentaires quant à sa stabilité, notamment dans des contextes où les données sont déséquilibrées.

Bibliographie

Leo Breiman.(2001), Random forests.,*Machine learning*, 45(1):5–32.

Gérard Biau.(2012), Analysis of a Random Forests Model.,*Journal of Machine Learning Research* , 13 (2012) 1063-1095.

Chao Chen, Andy Liaw, Leo Breiman, et al.(2004), Using random forest to learn imbalanced data.,*University of California, Berkeley*, 10(1-12):24.

Jason Brownlee. (2020), *Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning.*, Machine Learning Mastery.

Robert O'Brien and Hemant Ishwaran.(2019), A random forests quantile classifier for class imbalanced data.,*Pattern recognition*, 90:232–249.

Saïp Ciss. (2014), *Forêts uniformément aléatoires et détection des irrégularités aux cotisations sociales.*, PhD thesis, Paris 10.

Stefan Wager and Susan Athey. (2018), Estimation and inference of heterogeneous treatment effects using random forests., *Journal of the American Statistical Association* , 113 (523):1228–1242.

Sylvain Arlot and Robin Genuer.(2014), Analysis of purely random forests bias.,*arXiv preprint* , arXiv:1407.3939.

Wei Peng, Tim Coleman, and Lucas Mentch. (2022), ates of convergence for random forests via generalized u-statistics, *Electronic Journal of Statistics*, 16(1):232–292.

Gestion de données mixtes dans le cadre de données déséquilibrées : application au scoring de clients

Abdoulaye SAKHO[1] & Emmanuel MALHERBE[2] & Erwan SCORNET [3]

[1] *Atefact et LPSM, France aabdoulaye.sakho@artefact.com*
[2] *Artefact,, France et emmanuel.malherbe@artefact.com*
[3] *LPSM, France et erwan.scornet@polytechnique.edu*

Résumé. Cette étude examine les classificateurs à base d'arbres pour la détection d'événements rares dans le cadre de la classification binaire, en se concentrant sur les données tabulaires. Notre analyse est basée sur plusieurs algorithmes fondamentaux développés pour traiter le déséquilibre des classes, en particulier les stratégies qui génèrent des échantillons synthétiques dans la classe minoritaire. Cependant, peu de recherches ont exploré en profondeur la régénération des distributions des classes minoritaires lorsqu'il s'agit de caractéristiques mixtes composées à la fois de variables continues et catégorielles. Nous avons démontré l'importance de générer de nouveaux échantillons synthétiques qui présentent une combinaison de caractéristiques catégorielles déjà présentes dans les échantillons originaux. Nous montrons que le non-respect de cette propriété entraîne une baisse des performances prédictives. Nous proposons une nouvelle stratégie de suréchantillonnage, MGS+DRF, adaptée aux caractéristiques mixtes et préservant la cohérence des caractéristiques catégorielles. La méthode proposée utilise un estimateur de densité à noyau pour générer les caractéristiques continues, tandis qu'un modèle basé sur un arbre de distribution dessine, à partir des échantillons originaux, la combinaison des caractéristiques catégorielles. Nous évaluons l'efficacité de MGS+DRF à l'aide de plusieurs ensembles de données open-source et les résultats sont prometteurs.

Mots-clés. Classification, Données déséquilibrées, SMOTE, Données tabulaires.

Abstract. This study investigates tree-based classifiers for detecting rare events within the binary classification framework, focusing on tabular data. Our analysis is based on several seminal algorithms developed to address class imbalance, in particular strategies that generate synthetic samples in the minority class. However, limited research has extensively explored the regeneration of minority class distributions when dealing with mixed features composed of both continuous and categorical variables. We proved the importance of generating new synthetic samples that have a combination of categorical features already present in the original samples. We show that not respecting this property leads to lower predictive performances. We propose a novel oversampling strategy, MGS+DRF, tailored for mixed features and preserving the coherence of categorical features. Our proposed method uses a kernel density estimator to generate the continuous features while a distributional tree-based model draw, from original samples, the combination of categorical features. We evaluate MGS+DRF effectiveness using multiple open-source data sets and the results are promising.

Keywords. Classification, Imbalanced data, SMOTE, Tabular data.

1 Introduction

Le traitement du déséquilibre des classes dans les tâches de classification binaire représente un défi important pour les praticiens dans diverses applications d'apprentissage automatique [Khalilia et al., 2011, Nguyen and Duong, 2021], telles que la détection de la fraude [Hassan and Abraham, 2016, Li et al., 2024]. Dans des secteurs comme la banque, il s'agit d'un problème majeur. D'une part, la grande majorité des clients effectuent des transactions légitimes. D'autre part, les quelques transactions frauduleuses ont un impact important sur les opérations et la réputation.

Plusieurs travaux précurseurs ont introduit des stratégies de rééquilibrage afin d'améliorer les performances prédictives du classificateur sélectionné [voir Chawla et al., 2002, Mukherjee and Khushi, 2021]]. Ces stratégies peuvent être divisées en deux catégories [see Sakho et al., 2024], au niveau du modèle et au niveau des données. Les stratégies de rééquilibrage au niveau du modèle, telles que la pondération des classes, adaptent ou modifient un hyperparamètre d'un algorithme existant pour traiter le déséquilibre des données. Les approches au niveau des données agissent principalement sur l'ensemble de données d'origine en sur-échantillonnant ou en sous-échantillonnant les données. Le présent document se concentre sur les procédures au niveau des données et plus particulièrement sur les procédures synthétiques, qui peuvent être décrites comme des stratégies générant de nouveaux échantillons synthétiques dans la classe minoritaire. Les travaux antérieurs sur les ensembles de données déséquilibrés ont conduit à l'introduction de nombreuses stratégies de rééquilibrage [Chawla et al., 2002, He et al., 2008, Han et al., 2005]. Cependant, la plupart des stratégies sont conçues pour traiter des ensembles de données avec des caractéristiques numériques et la grande majorité des données sont déséquilibrées.

Les stratégies capables de traiter les caractéristiques catégorielles sont essentielles dans un cadre de données déséquilibrées (par exemple, la catégorie d'emploi, le pays), car les caractéristiques catégorielles ont autant d'impact sur la décision finale que les caractéristiques continues. Enfin, les caractéristiques catégorielles peuvent être divisées en caractéristiques nominales et ordinales. Cette distinction dépend du problème donné et de l'utilisation des caractéristiques catégorielles données par les praticiens.

En mettant l'accent sur les stratégies de rééquilibrage synthétique, cet article traite du cadre de classification binaire sur des ensembles de données tabulaires dans ce qui suit.

Motivations Notre principal objectif est de veiller à ce que les caractéristiques synthétiques générées soient cohérentes. Nous considérons que les échantillons synthétiques sont cohérents s'ils peuvent être jugés comme des échantillons plausibles par le propriétaire de l'entreprise. La cohérence des caractéristiques catégorielles est cruciale dans plusieurs applications d'apprentissage automatique, telles que la détection des fraudes. En effet, nous ne voulons pas introduire d'échantillons non cohérents qui peuvent rendre notre modèle moins efficace, moins compréhensible pour l'entreprise et augmenter les risques du modèle.

2 Méthode proposée

Au sein de cette section, nous présentons la méthodologie que nous proposons pour traiter un mélange de données continues et catégorielles.

Notations Soit $\{x_i, y_i\}_{i=1...N}$ les N paires indépendantes et identiquement distribuées comme (X, Y). X est une variable aléatoire qui prend des valeurs dans $\mathcal{X} \subset \mathbb{R}^d$ et Y $dans\{0, 1\}$. Nous supposons que n échantillons appartiennent à la classe dont l'étiquette est $Y = 1$. En outre, nous supposons que $n << N - n$, ce qui entraîne un déséquilibre entre la classe minoritaire $Y = 1$ et la classe majoritaire $Y = 0$. Nous désignons par f la distribution des caractéristiques continues conditionnées à la classe minoritaire, c'est-à-dire $f(x) = P(X = x|Y = 1)$.

Nous rappelons que notre principal compétiteur SMOTE-NC rélise une interpolation ;linéaire pour gérer les variables continues dans un premier temps. Ensuite, SMOTE-NC effectue un vote variable catégorielle par variable catégorielle pour générer la nouvelle combinaison de variables catégorielles.

2.1 Gestion des variables continues

Sakho et al. [2024] a présenté une stratégie de rééquilibrage basée sur les smotes, appelée Multivariate Gaussian SMOTE (MGS). Elle utilise des distributions gaussiennes pour générer de nouveaux échantillons synthétiques. En effet, une fois le point central sélectionné, ses K plus proches voisins sont utilisés pour estimer la moyenne empirique et la matrice de covariance d'une distribution gaussienne. L'échantillon généré est alors tiré de cette distribution gaussienne. Cette opération est effectuée pour chaque point central. Une itération de la procédure MGS est décrite dans Algorithm 1.

Algorithm 1 Multivariate Gaussian SMOTE iteration.

Input: Minority class samples $X_1, \ldots, X_n$, number K of nearest-neighbors.
Select uniformly X_c among $X_1, \ldots, X_n$.
Denote by I the set of the $K + 1$ nearest-neighbors of X_c among $X_1, \ldots, X_n$ including X_c (w.r.t. L_2 norm).
$\hat{\mu} \leftarrow \frac{1}{K+1} \sum_{x \in I} x$
$\hat{\Sigma} \leftarrow \frac{1}{K+1} \sum_{x \in I} (x - \hat{\mu})^T (x - \hat{\mu})$
Sample $Z \sim \mathcal{N}\left(\hat{\mu}, \hat{\Sigma}\right)$
Return Z

GS peut être interprété comme un estimateur de densité par noyau à largeur de bande adaptative avec un noyau gaussien. Plus précisément, il appartient à la famille des estimateurs de lissage d'échantillon. Pour x $dans\mathcal{X}$, la densité $\hat{f}(x)$ de notre estimateur est

de

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}|H_i|} \exp\left(-\frac{1}{2}(x-\hat{\mu_i})^T\hat{\Sigma}_i^{-1}(x-\hat{\mu_i})\right), \tag{1}$$

où I est l'ensemble des $K+1$ plus proches voisins de x_i parmi $x_1, \ldots, x_n$ incluant x_i (par rapport à la norme L_2). $\hat{\mu_i}$ et $\hat{\Sigma}_i$ sont estimés à partir du plus proche voisin pour chaque point central.

Nous remarquons que si $K = n$, alors la stratégie de suréchantillonnage du noyau gaussien est égale à ROSE [see Menardi and Torelli, 2014]. En outre, si les bandes passantes sont pondérées, nous tombons dans le cas introduit par Tang and He [2015].

MGS présente de meilleures performances prédictives que SMOTE-NC lorsque les caractéristiques d'entrée sont continues [see Sakho et al., 2024]. En outre, Stocksieker et al. montre que SMOTE peut être considéré comme un estimateur de densité de noyau avec un noyau uniforme.

Les matrices de covariance des MGS peuvent être estimées à l'aide de méthodes de shrinkage telles que celles de Ledoit-Wolf ou de l'OAS [voir, Ledoit and Wolf, 2004, Chen et al., 2010, pour plus de détails]. Ce choix a une incidence sur les performances de l'algorithme proposé dans un contexte de haute dimension.

2.2 Méthode finale proposée

Nous présentons MGS+`MultiOutputClassifier`, qui est l'une de nos nouvelles procédures pour générer de nouveaux échantillons synthétiques lorsque nous traitons des ensembles de données contenant à la fois des caractéristiques continues et catégorielles. Dans cette procédure, un `MultiOutputClassifier` est formé sur toutes les caractéristiques continues des échantillons minoritaires avec toutes les caractéristiques catégorielles comme cible. Ensuite, nous appliquons MGS pour générer les caractéristiques continues des nouveaux échantillons synthétiques. Enfin, le `MultiOutputClassifier` formé est utilisé pour prédire les caractéristiques catégorielles sur la base des caractéristiques continues des nouveaux échantillons synthétiques.

Le `MultiOutputClassifier` peut être une forêt aléatoire, un estimateur du plus proche voisin, par exemple. Nous incluons également dans notre étude un type particulier de forêt, la forêt aléatoire distributionnelle, décrite ci-dessous.

Distributional Random Forest [Cevid et al., 2022] is a machine learning algorithm that can derive the multivariate conditional probability $P(Y|X = x)$. DRF is a generalization of random forest [Breiman, 2001] to the multivariate case using Maximal Mean Discrepancy (MMD) as impurity criterion for regression framework. It is also possible with DRF to used $P(Y|X = x)$ with the induced forest weights in order to get an estimation target $\tau(x)$ (covariance matrix for example). The main difference between DRF and random forests is the fact that DRF associate weights to each sample that have been seen during fitting procedure. These weights are based on the occurrence of the sample in the leaf of the trees forest. During the predict procedure, probabilities are computed based on the occurrence of each training

sample to fall in the same leaf as the predicted sample for each tree. These probabilities for each training samples are then used in order to draw which of the training sample target will be returned by the distributional forest as prediction. Our proposed methods are available at `https://github.com/artefactory/smote_strategies_study_categorical`.

Algorithm 2 Single predict procedure of distributional forest

Input: T trees $\mathcal{T}_1, \ldots, \mathcal{T}_T$, a new unlabeled sample x_{new}, $\mathcal{L}_k(x_{new})$ the set of training samples which end up in the same leaf as x_{new} in the tree $\mathcal{T}_k$.
for $i \in [1, \ldots, n]$ **do**
$w_{x_{new}}(x_i) = \frac{1}{T} \sum_{k=1}^{T} \frac{\mathbb{1}_{\{x_i \in \mathcal{L}_k(x_{new})\}}}{|\mathcal{L}_k(x_{new})|}$
end for
Draw $\hat{y}$ from $\{y_1, \ldots, y_n\}$ based on $\{w_{x_{new}}(x_1), \ldots, w_{x_{new}}(x_n)\}$ probabilities.
Return $\hat{y}$

Motivations Dans la Section 2.1, nous avons expliqué que SMOTE et MGS peuvent être considérés comme des estimateurs de densité de noyau avec respectivement un noyau uniforme et un noyau gaussien. Nous pensons que MGS, avec son noyau gaussien, est plus adapté pour régénérer toute distribution sous-jacente de la classe minoritaire à partir d'un ensemble de données donné. En outre, Sakho et al. [2024] montre que MGS présente de bonnes performances prédictives par rapport à SMOTE sur divers ensembles de données.

Nous pensons donc que pour générer des caractéristiques catégorielles, nous ne devrions pas seulement baser notre choix sur le voisinage d'un point de données avec un vote caractéristique par caractéristique. Tout d'abord, parce que le plus proche voisin n'est pas pertinent dans un environnement à haute dimension. En outre, la prédiction de chaque caractéristique catégorielle une par une peut introduire des caractéristiques catégorielles non cohérentes dans les échantillons synthétiques. Par exemple, dans le secteur bancaire et dans certains pays, il n'est pas possible qu'un client non adulte (un enfant) soit marié. Avec la forêt distributionnelle, toutes les caractéristiques catégorielles sont prédites en même temps, ce qui préserve la cohérence des caractéristiques. Nous remarquons que SMOTE-NC peut générer des échantillons non cohérents.

Dans les sous-sections suivantes, nous décrivons nos lignes de base avant de présenter des simulations pour illustrer l'impact sur les performances prédictives de la génération d'échantillons synthétiques non cohérents.

3 Résultats sur données simulées

Nous générons un ensemble de données pour une tâche de classification binaire avec la classe 0 surreprésentée par rapport à la classe 1. En outre, seule une certaine combinaison de caractéristiques catégorielles est présente dans la classe minoritaire. Notre objectif est de détecter les procédures synthétiques qui génèrent des échantillons non cohérents et de déterminer si cela a un impact sur les performances prédictives.

A partir de notre protocole simulé, nous montrons que les stratégies qui créent de nouvelles combinaisons de caractéristiques catégorielles dans la classe minoritaire (c'est-à-dire celles avec $NCS > 0\%$) détériorent les performances prédictives du classificateur final. En effet, dans Table 1, nous remarquons que les AUC PR de SMOTE-NC, MGS+5-nn et MGS-NC sont inférieures à celles de la stratégie None.

Nous soulignons que l'importance de MGS+1nn et MGS+5nn ne diffère que du nombre de plus proches voisins utilisés pour le vote des caractéristiques catégorielles. Ainsi, le MGS+1nn est cohérent alors que le MGS+5nn ne l'est potentiellement pas. Dans Table 1, MGS+5nn réalise des performances inférieures à celles de MGS+1nn, ce qui souligne le caractère crucial de la cohérence pour les caractéristiques catégorielles. En outre, nous observons que MGS+DRF a réalisé les meilleures performances prédictives en termes d'AUC PR et d'AUC ROC. Ainsi, la méthode MGS+DRF que nous proposons est prometteuse.

Table 1: LightGBM trained on simulated data from the experimental protocol.

Strategy	None	CW	ROS	RUS	SMOTE-NC	MGS +1-nn	MGS +5-nn	MGS-NC	MGS +DRF
PR AUC	0.179	0.208	0.181	0.062	0.173	0.509	0.193	0.220	**0.513**
std	±0.074	±0.101	±0.086	±0.018	±0.072	±0.142	±0.072	±0.088	±0.144
ROC AUC	0.806	0.815	0.802	0.672	0.757	0.939	0.804	0.814	**0.941**
std	±0.044	±0.054	±0.049	±0.035	±0.044	±0.032	±0.039	±0.041	±0.028
$NCSO$	0%	0%	0%	0%	38%	0%	40%	38%	0%
std	±0	±0	±0	±0	±4	±0	±2	±4	±0

Une dernière observation dans Table 1 est que MGS-NC conduit à de meilleures performances prédictives que SMOTE-NC, alors que les deux stratégies diffèrent dans la manière dont les caractéristiques continues sont traitées. Ainsi, la densité du noyau gaussien de MGS semble mieux régénérer les distributions de la classe minoritaire que SMOTE, ce qui confirme les conclusions de Sakho et al. [2024].

4 Résultats sur données réelles

Sur les ensembles de données open-source, nous observons que notre méthode proposée MGS+DRF réalise la meilleure amélioration des performances prédictives pour les ensembles de données BankChurners et BankMarketing dans Tables 2 and 3. Ces améliorations des performances prédictives sont valables pour les mesures PR AUC, ROC AUC et precions@recall, ce qui est prometteur pour notre méthode.

D'autre part, nous remarquons que SMOTE-NC a des performances prédictives inférieures à celles de toutes les autres stratégies synthétiques au niveau des données (MGS+1-nn, MGS+5-nn, MGS-NC, MGS+DRF). Nous rappelons que SMOTE-NC est la stratégie synthétique de pointe pour les praticiens. Une autre observation concernant SMOTE-NC est que ses performances prédictives sont inférieures à celles de MGS-NC. Ainsi, sur des ensembles de

Table 2: LightGBM trained on BankChurners data set for different rebalancing strategies.

Strategy	None	CW	ROS	RUS	SMOTE-NC	MGS +1-nn	MGS +5-nn	MGS-NC	MGS +DRF
P@recall=0.2	0.894	0.870	0.847	0.632	0.850	0.913	0.910	0.908	**0.928**
std	±0.051	±0.056	±0.061	±0.123	±0.053	±0.050	±0.048	±0.057	±0.045
P@recall=0.5	0.623	0.614	0.570	0.316	0.583	0.682	0.661	0.680	**0.695**
std	±0.051	±0.050	±0.060	±0.057	±0.035	±0.047	±0.050	±0.059	±0.045
PR AUC	0.622	0.608	0.576	0.394	0.595	0.663	0.653	0.655	**0.668**
std	±0.024	±0.026	±0.024	±0.043	±0.021	±0.022	±0.024	±0.026	±0.023
ROC AUC	0.977	0.971	0.963	0.941	0.975	0.984	0.983	0.983	**0.984**
std	±0.005	±0.005	±0.006	±0.008	±0.004	±0.003	±0.003	±0.002	±0.002

Table 3: LightGBM trained on BankMarketing data set for different rebalancing strategies.

Strategy	None	CW	ROS	RUS	SMOTE-NC	MGS +1-nn	MGS +5-nn	MGS-NC	MGS +DRF
P@recall=0.2	0.119	0.118	0.115	0.106	0.093	0.128	0.126	0.128	**0.131**
std	±0.009	±0.009	±0.011	±0.012	±0.007	±0.012	±0.008	±0.014	±0.010
P@recall=0.5	0.084	0.079	0.078	0.071	0.063	0.093	0.092	0.090	**0.094**
std	±0.006	±0.005	±0.004	±0.004	±0.003	±0.003	±0.005	±0.003	±0.004
PR AUC	0.092	0.090	0.090	0.082	0.076	0.098	0.099	0.099	**0.101**
std	±0.008	±0.005	±0.006	±0.006	±0.004	±0.006	±0.005	±0.005	±0.005
ROC AUC	0.890	0.882	0.878	0.881	0.861	**0.899**	**0.899**	**0.899**	**0.899**
std	±0.003	±0.004	±0.003	±0.004	±0.004	±0.003	±0.002	±0.003	±0.003

données réelles, nous arrivons aux mêmes conclusions que dans Section 3 : l'estimateur à noyau gaussien MGS conduit à de meilleures performances prédictives que SMOTE.

En outre, dans Tables 2 and 3, nous remarquons que MGS+5-nn est associé à des performances prédictives plus faibles pour toutes les mesures que MGS+1-nn. Cette observation reflète à nouveau notre conclusion selon laquelle les résultats de Section 3, qui sont que la cohérence des caractéristiques catégorielles doivent être préservés.

References

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Domagoj Cevid, Loris Michel, Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

Yilun Chen, Ami Wiesel, Yonina C Eldar, and Alfred O Hero. Shrinkage algorithms for mmse covariance estimation. *IEEE transactions on signal processing*, 58(10):5016–5029, 2010.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

Amira Kamil Ibrahim Hassan and Ajith Abraham. Modeling insurance fraud detection using imbalanced data classification. In *Advances in Nature and Biologically Inspired Computing: Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing (NaBIC2015) in Pietermaritzburg, South Africa, held December 01-03, 2015*, pages 117–127. Springer, 2016.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee, 2008.

Mohammed Khalilia, Sounak Chakraborty, and Mihail Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, 11:1–13, 2011.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Kaidi Li, Tianmeng Yang, Min Zhou, Jiahao Meng, Shendi Wang, Yihui Wu, Boshuai Tan, Hu Song, Lujia Pan, Fan Yu, et al. Sefraud: Graph-based self-explainable fraud detection via interpretative mask learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5329–5338, 2024.

Giovanna Menardi and Nicola Torelli. Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, 28:92–122, 2014.

Mimi Mukherjee and Matloob Khushi. Smote-enc: A novel smote-based method to generate synthetic data for nominal and continuous features. *Applied system innovation*, 4(1):18, 2021.

Nam N Nguyen and Anh T Duong. Comparison of two main approaches for handling imbalanced data in churn prediction problem. *Journal of advances in information technology*, 12(1), 2021.

Abdoulaye Sakho, Emmanuel Malherbe, and Erwan Scornet. Do we need rebalancing strategies? a theoretical and empirical study around smote and its variants. https://hal.science/hal-04438941v4/file/smote.pdf, 2024.

Samuel Stocksieker, Denys Pommeret, and Arthur Charpentier. Generalized oversampling for learning from imbalanced datasets and associated theory: Application in regression.

Bo Tang and Haibo He. Kerneladasyn: Kernel based adaptive synthetic data generation for imbalanced learning. In *2015 IEEE congress on evolutionary computation (CEC)*, pages 664–671. IEEE, 2015.

Data Augmentation with Variational Autoencoder for Imbalanced Dataset

Samuel Stocksieker [1] & Denys Pommeret [1]

[1] *I2M, Aix-Marseille Université*

Résumé. L'apprentissage à partir d'une distribution déséquilibrée constitue un défi majeur en modélisation prédictive, car il entraîne généralement une diminution des performances des algorithmes standards. Différentes approches existent pour pallier ce problème, mais la majorité d'entre elles concernent les problèmes de classification, tandis que la régression reste peu explorée. Dans cet article, nous introduisons une nouvelle méthode visant à améliorer l'apprentissage sur des données tabulaires dans le cadre de la Imbalanced Regression (IR), qui demeure un enjeu important. Nous proposons d'utiliser les autoencodeurs variationnels (VAE), reconnus pour leur capacité à générer des données synthétiques et à capturer des représentations latentes de distributions complexes. Cependant, les VAE peuvent s'avérer inefficaces face aux défis posés par l'IR. Ainsi, nous développons une approche novatrice combinant les VAE avec un smoothed bootstrap, spécialement conçue pour traiter ces défis. Nous évaluons numériquement l'efficacité de cette méthode en la comparant aux approches concurrentes à travers des simulations et des ensembles de données présentant des déséquilibres en régression.

Mots-clés. Imbalanced, Regression, Variational Autoencoder, Representation Learning

Abstract. Learning from an imbalanced distribution presents a major challenge in predictive modeling, as it generally leads to a reduction in the performance of standard algorithms. Various approaches exist to address this issue, but many of them concern classification problems, with a limited focus on regression. In this paper, we introduce a novel method aimed at enhancing learning on tabular data in the Imbalanced Regression (IR) framework, which remains a significant problem. We propose to use variational autoencoders (VAE) which are known as a powerful tool for synthetic data generation, offering an interesting approach to modeling and capturing latent representations of complex distributions. However, VAEs can be inefficient when dealing with IR. Therefore, we develop a novel approach for generating data, combining VAE with a smoothed bootstrap, specifically designed to address the challenges of IR. We numerically investigate the scope of this method by comparing it against its competitors on simulations and datasets known for IR.

Keywords. Imbalanced, Regression, Variational Autoencoder, Representation Learning

1 Introduction

Les techniques de régression sont largement utilisées dans divers domaines tels que la finance, l'économie, la médecine et l'ingénierie afin de modéliser les relations entre les variables ex-

plicatives et une variable cible continue. Cependant, dans les applications réelles, les jeux de données présentent fréquemment un déséquilibre significatif dans la distribution de la variable cible, ce qui pose des défis spécifiques aux méthodes de modélisation traditionnelles ((Krawczyk, 2016), (Fernández et al., 2018a)). En effet, les approches standard considèrent généralement toutes les valeurs comme également importantes et sont entraînées à minimiser l'erreur moyenne. Cependant, les valeurs rares et/ou extrêmes étant peu nombreuses dans l'ensemble d'apprentissage, elles ont une influence marginale sur la minimisation de l'erreur au cours du processus d'apprentissage. Pourtant, ces valeurs sont souvent d'une grande importance.

Cette problématique a été largement étudiée dans le cadre de la classification (Haixiang et al., 2017), (Fernández et al., 2018b). Cependant, très peu de travaux ont abordé la question de la régression déséquilibrée (Imbalanced Regression - IR), malgré son importance dans de nombreuses applications réelles (Krawczyk, 2016). La régression pose en effet des défis supplémentaires par rapport à la classification : (i) les valeurs rares ou minoritaires ne sont pas directement identifiables en raison de la nature continue de la distribution, (ii) le degré de déséquilibre est difficilement mesurable, (iii) l'ampleur du rééquilibrage est complexe à évaluer et (iv) la génération de données synthétiques nécessite également de produire de nouvelles valeurs pour la variable cible.

Dans le cas des données tabulaires, les solutions de prétraitement, notamment la génération de données synthétiques, sont souvent privilégiées (Branco et al., 2016b; He and Ma, 2013) en raison de leur universalité, car elles permettent d'utiliser n'importe quelle technique de régression standard. Les VAE sont des modèles de génération de données qui utilisent un réseau de neurones pour encoder les observations dans un espace latent continu, où de nouvelles données peuvent être générées de manière probabiliste. Cette approche s'est révélée particulièrement efficace pour la génération de données non structurées, notamment grâce à sa capacité à capturer des relations non linéaires.

Notre objectif est ici de générer des valeurs rares artificielles tout en prenant en compte les corrélations observées. En effet, contrairement au cas de la classification, la régression impose de générer une nouvelle valeur pour la variable cible. Nous proposons donc d'utiliser une approche basée sur les VAE pour construire une nouvelle représentation latente des données, permettant une génération plus pertinente que celle obtenue à partir de l'espace initial des données. Dans cet article, nous proposons une modification simple et efficace des VAE afin d'adapter leur entraînement et leur génération au problème de la régression déséquilibrée sur des données structurées. Nos principales contributions peuvent être résumées comme suit :

- Proposition d'une nouvelle fonction de perte adaptée à l'entraînement des VAE dans le cadre de la régression déséquilibrée.
- Élaboration d'un nouveau schéma de pondération permettant de contrôler l'échantillonnage des valeurs rares lorsque la variable cible est continue.
- Introduction d'un nouveau générateur de données synthétiques combinant l'apprentissage des VAE et une méthode de génération non paramétrique : le bootstrap lissé. Cette technique permet de générer des données à partir d'une observation "graine" en utilisant son voisinage.

2 DAVID: A Novel Synthetic Data Generator

Soit $\boldsymbol{x} = (x_{ij})_{i=1,\cdot,n;j=1,\cdot,p} \in \mathcal{X} \subset \mathbb{R}^{n\times p}$ un jeu de données composé de n observations et p variables, où x_{ij} représente la variable j pour l'observation i. Soit $y = (y_i)_{i=1,\cdot,n} \in \mathcal{Y} \subset \mathbb{R}^{\kappa}$ la variable cible associée, qui est déséquilibrée et continue.

Les autoencodeurs permettent de construire un espace de représentation latente des données, capturant ainsi les relations non linéaires entre les variables. En introduisant une composante stochastique dans l'espace latent, les autoencodeurs variationnels (VAE) entraînent des variables latentes aléatoires représentant les données et génèrent naturellement de nouvelles données synthétiques très proches des données initiales. Les VAE sont conçus pour reconstruire les entrées à partir des variables latentes. Cependant, dans un cadre de régression, la variable cible Y ne doit pas être traitée comme une simple variable explicative X. Nous proposons donc de construire un VAE mixte : celui-ci prend en entrée à la fois les variables explicatives X et la variable cible Y et cherche à les reconstruire, mais avec une pondération spécifique appliquée à Y. La fonction de perte globale est donc classiquement définie comme suit :

$$\mathcal{L}(\theta, \phi, \boldsymbol{x}, y) = \beta_x \mathbb{E}_q[log\, p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta_{KL} D_{KL}(q_\theta(z|\boldsymbol{x})||p_\theta(\boldsymbol{z})) + \beta_y \mathbb{E}_q[log\, p_\theta(y|\boldsymbol{z})]$$

où $\{\beta_x,\ \beta_y,\ \beta_{KL}\}$ sont les poids associés, $\mathbb{E}_q[log\, p_\theta(\boldsymbol{x}|\boldsymbol{z})]$ (resp. $\mathbb{E}_q[log\, p_\theta(y|\boldsymbol{z})]$) représente la fonction de perte de reconstruction pour $\boldsymbol{x}$ (resp. y) et $D_{KL}(q(z|x), p(z|x))$ représente la divergence de Kullback-Leibler pour la régularisation. Plus précisément, $q_\theta(z|\boldsymbol{x})$ représente la distribution des variables latentes z conditionnée aux données d'entrée $\boldsymbol{x}$ et $p_\theta(\boldsymbol{z})$ représente la distribution a priori, souvent choisie comme une distribution gaussienne. L'avantage d'utiliser un VAE dans ce contexte est d'assurer la régularité (continuité de la distribution a priori) de l'espace latent grâce à une pénalisation, permettant ainsi une génération cohérente des données.

Il a été démontré que la fonction de perte d'erreur quadratique moyenne (Mean Squared Error, MSE) n'est pas efficace pour traiter le problème de la régression déséquilibrée (Ren et al., 2022). En effet, elle favorise mécaniquement les valeurs fréquentes, car celles-ci réduisent significativement la fonction de perte, ce qui conduit à négliger les valeurs rares. Nous proposons ici l'utilisation d'une fonction de perte équilibrée prenant en compte la fréquence de la variable cible Y. L'objectif est d'entraîner un modèle capable de générer de nouvelles données synthétiques pour les valeurs rares de Y.

Dans les tâches de classification, les valeurs rares sont directement identifiables, car elles appartiennent à une ou plusieurs classes minoritaires. Cependant, en régression, où le support de la variable cible est continu, la définition même des valeurs rares constitue un premier défi. Pour y remédier, nous proposons de pondérer les observations lors de la phase d'entraînement en utilisant l'inverse de la densité empirique de Y. Autrement dit, plus une valeur est rare, plus son poids sera élevé, et inversement. Enfin, nous suggérons d'atténuer l'effet de cette pondération à l'aide d'un paramètre α. Plus ce paramètre est élevé, plus la différence entre les poids sera marquée. La fonction de perte globale devient ainsi :

$$\mathcal{L}(\theta, \phi, \boldsymbol{x}, y) = \beta_x \mathbb{E}_q[log\, p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta_{KL} D_{KL}(q_\theta(z|\boldsymbol{x})||p_\theta(\boldsymbol{z})) + \frac{\beta_y}{\widehat{f}(Y)^\alpha} \mathbb{E}_q[log\, p_\theta(y|\boldsymbol{z})] \quad (1)$$

Étant donné que la densité de Y est inconnue, nous proposons de l'estimer à l'aide d'un estimateur à noyau (avec un paramètre de lissage garantissant sa convergence, tel que la matrice de bande passante de Silverman (Silverman, 1986) ou celle de Scott (2015)).

Comme mentionné précédemment, l'objectif et l'avantage des Variational Autoencoders (VAEs) résident dans leur capacité à générer des données similaires à celles sur lesquelles le modèle a été entraîné. Typiquement, un VAE classique (utilisant des distributions gaussiennes) cherche à calibrer les paramètres μ_i et σ_i d'une distribution normale latente à partir de l'observation x_i.

Le principal défi dans la Imbalanced Regression (IR) est que les algorithmes ont du mal à modéliser les valeurs rares. Les autoencodeurs classiques rencontrent également ce problème. Dans la section précédente, nous avons proposé un premier ajustement sous la forme d'une fonction de perte MSE équilibrée (balanced MSE), afin d'améliorer le processus d'apprentissage du VAE. Cependant, la variance σ_i attribuée aux valeurs rares pourrait rester importante, ce qui pourrait conduire à une génération de données qui ne reflète pas fidèlement les données d'origine. En effet, le VAE construit une représentation latente propre à chaque observation, c'est-à-dire qu'il calibre les paramètres de chaque variable latente de manière indépendante : les représentations latentes sont supposées i.i.d. (independent and identically distributed).

Il est crucial de souligner que la génération de données concerne principalement les valeurs rares. L'objectif est de construire un échantillon d'entraînement composé à la fois de valeurs "majoritaires" observées et de valeurs rares, qu'elles soient observées ou générées.

Nous proposons ici un second niveau de traitement pour améliorer la génération des données rares. Tout d'abord, nous définissons les poids d'importance utilisés pour l'échantillonnage afin de rééquilibrer le processus d'apprentissage et ainsi permettre une meilleure modélisation des valeurs rares :

$$\omega_i := \frac{1}{\hat{f}_Y(y_i)^\alpha} \tag{2}$$

Ensuite, plutôt que de générer les nouvelles données de manière conventionnelle avec le VAE, nous suggérons d'utiliser un bootstrap lissé (smoothed bootstrap, voir (Silverman and Young, 1987), (Hall et al., 1989), (De Angelis and Young, 1992)) appliqué à la matrice μ de dimensions $n \times q$, qui contient les valeurs μ_{ij} représentant la moyenne des nouvelles représentations de X dans l'espace latent de dimension q :

$$\mu := \{\mu_{ij}, i = 1, \cdots, n\,;\, j =, \cdots, q\} \tag{3}$$

Le smoothed bootstrap (SB) consiste à tirer des échantillons à partir d'estimateurs à noyaux de la densité de probabilité sous-jacente. Il peut être décomposé en deux étapes : tout d'abord, une observation seed est tirée aléatoirement, puis un bruit aléatoire issu de l'estimateur à noyaux est ajouté afin d'obtenir un nouvel échantillon. Ici, la première étape est représentée par le poids d'échantillonnage ω_i et la seconde par une génération à l'aide d'un noyau appliqué à μ.

Les propriétés de convergence du smoothed bootstrap ont été étudiées dans (De Martini and Rapallo, 2008) et (Falk and Reiss, 1989). Nous proposons d'utiliser un mélange de noyaux multivariés (par exemple, Gaussiens) pour générer des données synthétiques :

$$g_{Z^*}(z^*|\mu) = \sum_{i=1,\cdots,n} \omega_i K_i(z^*, \mu),$$

où $(K_i)_{i=1,\cdots,n}$ est une collection de noyaux et $(\omega_i)_{i=1,\cdots,n}$ est une séquence de poids positifs avec $\sum_{i=1,\cdots,n} \omega_i = 1$. L'indice * indique ici les données synthétiques générées.

L'un des principaux avantages du smoothed bootstrap est qu'il ne nécessite aucun paramètre supplémentaire. Cette méthode repose sur l'estimation de la densité par noyaux, pour laquelle l'estimation du paramètre de lissage H est déjà optimisée. En effet, (Scott, 2015) et (Silverman, 1986) proposent des paramétrisations optimales de la matrice de lissage afin d'assurer la cohérence de l'estimateur. De plus, il est important de noter que le VAE génère un espace latent régulier, offrant ainsi un cadre idéal pour l'application du smoothed bootstrap. Ce n'est pas nécessairement le cas de l'espace d'origine, qui peut présenter des discontinuités ou des distributions bornées. Enfin, à l'instar des algorithmes Gaussian Noise (Lee and Sauchi, 2000) et ROSE (Menardi and Torelli, 2014), nous introduisons un paramètre permettant de régler le niveau de bruit et de contrôler la génération. Ce paramètre multiplicatif (¡1) est appliqué directement à la matrice de lissage.

Dans un VAE classique, la variance σ_i associée aux valeurs rares pourrait être instable et donc impacter la qualité de génération de nouvelles valeurs. Nous proposons de remédier à ce problème avec notre approche DAVID, qui considère l'ensemble des μ_i et propose une fonction générative conjointe définie comme suit :

$$g_{z^*}(z^*|\mu) = \sum_{i=1}^{n} \omega_i K_{H_n}(z^* - \mu_i),$$

où K est un noyau Gaussien et $H_n := \eta \times Var(\mu)$ est défini par la règle de Silverman :

$$\eta = \left(\frac{4}{p+2}\right)^{\frac{1}{d+4}} n^{\frac{-1}{d+4}} \tag{4}$$

Ici, $Var(\mu)$ représente la matrice de variance de μ.

On observe que la génération de z^* à partir d'une seed μ_i est réalisée via un noyau centré sur celle-ci, en fonction de son voisinage et des caractéristiques de la distribution de μ. Contrairement aux VAEs classiques, qui génèrent à partir d'une distribution gaussienne paramétrée par (μ_i, σ_i), notre approche exploite le voisinage de la représentation latente des valeurs rares pour générer de nouvelles valeurs.

3 Experiments

Pour tester notre approche sur des ensembles de données réels, nous comparons les résultats sur des ensembles de données de référence pour la régression déséquilibrée (IR) issus de

(Branco et al., 2019)[1]. Pour comparer notre approche à l'état de l'art, nous simulons également des données synthétiques en utilisant les méthodes suivantes :

- Un Autoencodeur Variationnel Tabulaire ((Xu et al., 2019; Patki et al., 2016) : *TVAE*.
- Un Réseau Adversarial Génératif Tabulaire Conditionnel ((Xu et al., 2019; Patki et al., 2016) : *CTGAN*.
- Un Réseau Adversarial Génératif basé sur les Copules, ((Patki et al., 2016)) : *CopGAN*.
- Un suréchantillonnage appliqué sur l'ensemble d'entraînement, utilisant l'approche UBL ((Branco et al., 2016a; Wu et al., 2022)) : *ILRro*.
- La méthode SMOTE pour la Régression ((Torgo et al., 2013; Branco et al., 2016a)) : *ILRsmote*.
- L'injection de Bruit Gaussien pour la Régression ((Branco et al., 2017, 2016a)) : *ILRgn*.

Dataset	bank8FM	abalone	boston	NO2
Train	mean (std)	mean (std)	mean (std)	mean (std)
Baseline	1,87 (0,19)	6,84 (0,29)	44,44 (9,57)	0,48 (0,04)
OS	2,61 (0,4)	6,19 (0,17)	41,01 (11,07)	0,47 (0,06)
kTrain	3,3 (0,52)	5,81 (0,18)	37,49 (12,47)	0,46 (0,07)
DAVID (ours)	**1,68 (0,13)**	**5,52 (0,22)**	**34,86 (12,12)**	**0,4 (0,04)**
kPCA (ours)	3,35 (0,52)	5,83 (0,21)	39,22 (11,64)	0,46 (0,07)
kKPCA (ours)	2,57 (0,51)	5,81 (0,18)	38,99 (11,06)	0,44 (0,05)
TVAE	2,56 (0,65)	7,17 (0,33)	45,2 (12,59)	0,5 (0,06)
CTGAN	8,96 (4,55)	7,43 (0,32)	51,77 (15,11)	0,54 (0,05)
CopGAN	8 (3,08)	7,5 (0,3)	55,79 (12,37)	0,57 (0,08)
ILRro	3,37 (0,36)	6,42 (0,2)	46,11 (10,85)	0,48 (0,04)
ILRsmote	2,6 (0,33)	5,78 (0,2)	44,3 (9,19)	0,47 (0,04)
ILRgn	2,51 (0,38)	6,41 (0,18)	42,78 (10,73)	0,46 (0,04)

Table 1: Experiments results: wMSE for test prediction

Les résultats présentés dans le Tableau 1 confirment ceux de l'illustration : DAVID (*kBVAEw*) obtient de meilleures performances que l'échantillon d'entraînement initial (*baseline*)ainsi que les approches de l'état de l'art (*ILRro, ILRsmote, ILRgn*). Ces résultats indiquent qu'il est préférable de générer des données synthétiques à partir de l'espace latent plutôt qu'à partir de l'espace initial (*kTrain*: Smoothed Boostrap dans l'espace initial et *OS*: oversampling). Enfin, ces expériences montrent que l'utilisation de générateurs classiques de données synthétiques (*TVAE, CTGAN, copGAN*) n'est pas recommandée dans le cadre de la régression déséquilibrée. La même approche a été utilisée sur une analyse en composantes principales (*kPCA*) et son extension à noyau (*kKPCA*), mais les résultant sont plus faibles.

[1]Le dépôt dédié "Data Sets for Imbalanced Regression Learning" est disponible à l'adresse suivante : `https://paobranco.github.io/DataSets-IR/`

References

Branco, P., Ribeiro, R. P., and Torgo, L. (2016a). Ubl: an r package for utility-based learning. *arXiv preprint arXiv:1604.08079*.

Branco, P., Torgo, L., and Ribeiro, R. P. (2016b). A survey of predictive modeling on imbalanced domains. *ACM computing surveys (CSUR)*, 49(2):1–50.

Branco, P., Torgo, L., and Ribeiro, R. P. (2017). Smogn: a pre-processing approach for imbalanced regression. In *First international workshop on learning with imbalanced domains: Theory and applications*, pages 36–50. PMLR.

Branco, P., Torgo, L., and Ribeiro, R. P. (2019). Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing*, 343:76–99.

De Angelis, D. and Young, G. A. (1992). Smoothing the bootstrap. *International Statistical Review/Revue Internationale de Statistique*, pages 45–56.

De Martini, D. and Rapallo, F. (2008). On multivariate smoothed bootstrap consistency. *Journal of statistical planning and inference*, 138(6):1828–1835.

Falk, M. and Reiss, R.-D. (1989). Weak convergence of smoothed and nonsmoothed bootstrap quantile estimates. *The Annals of Probability*, pages 362–371.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018a). *Learning from imbalanced data sets*, volume 10. Springer.

Fernández, A., Garcia, S., Herrera, F., and Chawla, N. V. (2018b). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73:220–239.

Hall, P., DiCiccio, T. J., and Romano, J. P. (1989). On smoothing and the bootstrap. *The Annals of Statistics*, pages 692–704.

He, H. and Ma, Y. (2013). *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232.

Lee and Sauchi (2000). Noisy replication in skewed binary classification. *Computational Statistics and Data Analysis*, 34(2):165–191.

Menardi and Torelli (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122.

Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410.

Ren, J., Zhang, M., Yu, C., and Liu, Z. (2022). Balanced mse for imbalanced visual regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7926–7935.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Silverman, B. and Young, G. (1987). The bootstrap: to smooth or not to smooth? *Biometrika*, 74(3):469–479.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Torgo, L., Ribeiro, R. P., Pfahringer, B., and Branco, P. (2013). Smote for regression. In *Portuguese conference on artificial intelligence*, pages 378–389. Springer.

Wu, W., Kunz, N., and Branco, P. (2022). Imbalancedlearningregression-a python package to tackle the imbalanced regression problem. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 645–648. Springer.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*.

Biostatistique 1

Bayesian survival time prediction of French Childhood Cancer Survivors' Study using re-parameterized mixture distribution

Malek Ben Salah [1] and Kaniav Kamary [2]

[1] *Université Paris Saclay, Centrale Supélec, Laboratory of mathematics and informatics for complexity and systems, France malek.ben-salah@centralesupelec.fr*
[2] *ICJ, INSA de Lyon, Lyon, France*
kaniav.kamary@insa-lyon.fr

Résumé. Depuis ces cinquante dernières années, les progrès thérapeutiques ont considérablement améliorés le taux de survie du cancer pédiatrique. Malgré la croissance de taux de survie, il a été démontré que des traitements contre le cancer peuvent causer des complications préoccupantes à long terme, connues sous le nom d'effets tardifs, notamment des pathologies cardiaques telles que les valvulopathies. L'étude des personnes ayant survécu au cancer pédiatrique et l'analyse de survie ont donc largement fait l'objet de recherche scientifique pluridisiplinaire dans la littérature pour une meilleure prise en charge de l'après-cancer. Bien que les méthodes existantes d'analyse de survie, comme les modèles à risques proportionnels de Cox, se soient révélées précieuses, elles échouent souvent à capturer l'hétérogénéité inhérente aux populations de patients et les schémas complexes de survie caractéristiques des complications cardiaques post-cancer. Nous présentons un nouveau cadre statistique basé sur un modèle de mélange gaussien reparamétré qui répond à ces limitations tout en maintenant l'interprétabilité, caractéristique cruciale pour les applications cliniques. Notre approche s'appuie sur la stratégie de reparamétrisation de moyenne-variance, associée à un algorithme robuste de Metropolis-within-Gibbs qui assure une inférence stable. La flexibilité du modèle permet d'identifier des sous-groupes de risque distincts parmi les survivants du cancer pédiatrique tout en fournissant des expressions analytiques pour l'estimation des paramètres, et améliorant significativement l'efficacité computationnelle des méthodes existantes. Grâce à des garanties théoriques et à une validation empirique, nous démontrons la capacité de notre cadre à prédire précisément les complications cardiaques chez les survivants du cancer pédiatrique, ce qui permettrait potentiellement des protocoles de surveillance et des interventions plus ciblés. Cette avancée méthodologique représente un progrès significatif dans l'évaluation personnalisée des risques pour le suivi à long terme des survivants du cancer.

Mots-clés. Distribution de mélange, Temps de survie, Analyse de survie, Distribution a priori, Analyse bayésienne, Fonction de lien de régression linéaire généralisée

Abstract. Recent advances in pediatric cancer treatment have significantly improved survival rates, yet they have also revealed concerning long-term complications, particularly cardiac pathologies such as Valvular Heart Disease (VHD). While existing survival analysis methods like Cox proportional hazards models have proven valuable, they often fail to capture the inherent heterogeneity in patient populations and complex survival patterns

characteristic of post-cancer cardiac complications. We present a novel statistical framework based on a reparametrized Gaussian mixture model that addresses these limitations while maintaining interpretability, a crucial feature for clinical applications. Our approach leverages the location-scale reparametrization strategy, coupled with a robust Metropolis-within-Gibbs sampling algorithm that ensures stable inference settings. The model's flexibility allows it to identify distinct risk subgroups among childhood cancer survivors while providing closed-form expressions for parameter estimation, significantly improving upon existing methods' computational efficiency. Through theoretical guarantees and empirical validation, we demonstrate our framework's capability to more accurately predict cardiac complications in pediatric cancer survivors, potentially enabling more targeted monitoring protocols and interventions. This methodological advancement represents a significant step forward in personalized risk assessment for long-term cancer survivor care.

Keywords. Mixture distribution, Survival time, Prior time, Bayesian analysis, Generalized linear regression link function

1 Introduction and Literature Review

Progress in pediatric cancer treatment has improved survival rates over recent decades, leading to increased attention on long-term survivor outcomes. Following childhood cancer therapies such as radiotherapy and chemotherapy, cardiac pathology and Valvular Heart Disease (VHD) have emerged as significant health concerns for survivors (Chounta et al., 2023; Erdmann et al., 2021). The identification and prediction of heart disease occurrence in these survivors is essential for improving therapeutic interventions and post-treatment monitoring protocols. To effectively analyze and predict the onset of such cardiac complications, accurate modeling of time-to-event data is crucial (Armenian et al., 2018). The challenge of modeling time-to-event data has historically been addressed through classical frameworks such as the Cox proportional hazards and accelerated failure time (AFT) models (Cox, 1972; Wei, 1992). These methods assume homogeneous populations and impose restrictive parametric forms (e.g., Weibull or log-normal distributions). While interpretable, they struggle to capture multimodal or heterogeneous survival patterns (Van Ness et al., 2023; Zhang et al., 2024). These limitations, particularly in clinical settings where patient populations often exhibit distinct risk profiles, motivated the development of more flexible approaches.

To address unobserved heterogeneity, mixture models emerged as a powerful tool. Early work by Farewell (1982) introduced finite mixtures of parametric distributions (e.g., Weibull) to model subpopulations with differing survival mechanisms. Similarly, cure models framed survival data as mixtures of "cured" and "susceptible" populations, enabling inference on both incidence and latency (Maller et al., 2024). While these approaches improved flexibility over classical methods, they remained tethered to exponential-family distributions, limiting their ability to model asymmetric or complex hazards often present in real-world survival data.

The increasing complexity of survival models has highlighted the crucial trade-off between model flexibility and interpretability. In clinical settings particularly, understanding model predictions is often as important as their accuracy, since medical decisions require clear justification and transparent risk assessment. This has led to renewed interest in Gaussian mixture models (GMMs) for their unique combination of modeling power and mathematical tractability. GMMs can capture complex, multimodal distributions while maintaining the analytical advantages of normal distributions, including closed-form expressions for moments and marginals (Plate, 1999). However, their flexibility often comes at the cost of interpretability, as the relationship between model parameters and survival outcomes can be obscure.

Several researchers have proposed methods to balance accuracy and interpretability in these models (see for example Chen et al. (2019); Kunkel and Peruggia (2020); Safinianaini and Boström (2021)). Notable advances include the development of intrinsically interpretable GMMs by Alangari et al. (2023), which provide both global and local interpretations without compromising accuracy, or anchored Bayesian GMMs by Kunkel and Peruggia (2020), which assign meaningful labels to components during modeling.

The integration of covariates into mixture models represents another significant advancement, with numerous approaches proposed in recent years such as Gudicha and Vermunt (2013); Hoshikawa (2013); Frühwirth-Schnatter et al. (2019). The MoEClust suite of models allows covariates to influence both component weights and densities, demonstrating substantial improvements across various datasets (Murphy and Murphy, 2020). Papastamoulis and Perrakis (2024) introduced a Bayesian approach to clustering with covariates, advancing the field further. However, existing methods face several critical challenges: (i) assumptions of Gaussian-distributed covariates limit their applicability to real-world data, (ii) MCMC mixing remains problematic due to label switching and multimodal posterior distributions, and (iii) inference stability can be compromised by slow mixing rates, particularly in high-dimensional settings.

Our work addresses these fundamental limitations in three key ways. First, building on the reparametrization strategy of Kamary et al. (2018), we eliminate the need for distributional assumptions on covariates. Second, we derive closed-form expressions for sampling covariate parameters, substantially improving the stability of Gibbs sampling steps. This reformulation of the parameter space imposes principled constraints, resulting in improved MCMC mixing efficiency. Finally, our framework provides theoretical guarantees for convergence while maintaining computational efficiency. These advances collectively enable more reliable clustering and regression while preserving the flexibility that makes mixture models attractive for survival analysis applications.

2 Methodology

2.1 Modeling the survival time

Given a dataset $(X_i, T_i)_{i\in[\![1,n]\!]}$ where X_i are the covariates and T_i are the survival times. We model the r.v. $T \mid X$ as a mixture of log-normal distributions. This choice is motivated by the ability of this distribution to capture right-skewed heavy tailed survival data. We introduce a latent variable Z that for a given X indicates the component it belongs to.

$$\mathfrak{M}(T, X) : \begin{cases} Z \mid p_1, \dots, p_k \sim \text{Categorical}(p_1, \dots, p_k) \\ T \mid Z, X \sim \mathcal{LN}(\mu_Z(X), \sigma_Z^2(X)) \end{cases}$$

To make the model more tractable, we introduce the r.v. $Y := \log(T)$, so that Y follows a Gaussian mixture distribution. Thus, we move from the log-normal mixture in $\mathfrak{M}(T, X)$ to a normal mixture in $\mathfrak{M}(Y, X)$ defined by

$$\mathfrak{M}(Y, X) : \begin{cases} Z \mid p_1, \dots, p_k \sim \text{Categorical}(p_1, \dots, p_k) \\ Y \mid Z, X \sim \mathcal{N}(\mu_Z(X), \sigma_Z^2(X)) \end{cases} \tag{1}$$

We reparametrize the model (1) using the location-scale reparametrization of Kamary et al. (2018) where we denote by $\mu(X;\beta) = \mathbb{E}[Y|X] = \beta^\top X$ and $\sigma^2 = \mathbb{V}[Y|X]$.

$$\begin{cases} \mu_i = \mu(x_i;\beta) + \sigma\dfrac{\gamma_i}{\sqrt{p_i}}, \\ \sigma_i = \sigma\dfrac{\eta_i}{\sqrt{p_i}}, \end{cases} \quad \forall i \in [\![1,k]\!], \quad \text{s.t.} \quad \sum_{i=1}^{k}\sqrt{p_i}\gamma_i = 0, \quad \sum_{i=1}^{k}\left(\eta_i^2 + \gamma_i^2\right) = 1 \tag{2}$$

and the completed likelihood of the model is given by

$$\mathcal{L}(\Theta|\mathbf{x},\mathbf{y},\mathbf{z}) = \prod_{i=1}^{n}\prod_{j=1}^{k}\left[\frac{\sqrt{p_j}}{\sqrt{2\pi}\sigma\eta_i}\exp\left(-\frac{p_i(y_i - (\beta^\top x_i + \sigma\gamma_i/\sqrt{p_i}))^2}{2\sigma^2\eta_i^2}\right)\right]^{\delta_{z_i,j}} \tag{3}$$

Furthermore, mean and variance parameters in (2) can be decoupled by introducing a supplementary parameter, namely a radius φ such that

$$\sum_{i=1}^{k}\gamma_i^2 = \varphi^2, \quad \sum_{i=1}^{k}\eta_i^2 = 1 - \varphi^2.$$

Through this parametrization, $(\gamma_1, \dots, \gamma_k)$ lies at the intersection of a sphere of radius φ and a hyperplane of $\mathbb{R}^{k-1}$, while $(\eta_1, \dots, \eta_k)$ lies at the intersection of a sphere of radius $\sqrt{1-\varphi^2}$ and another hyperplane of $\mathbb{R}^{k-1}$. By further parametrizing these intersections, we introduce angular parameters $(\omega_i)_{i\in[\![1,k-2]\!]} \in [0,\pi]^{k-3} \times [0, 2\pi]$ and $(\xi_i)_{i\in[\![1,k-2]\!]} \in [0, \pi/2]^{k-1}$. This leads to the reparametrization $(\gamma_1, \dots, \gamma_k) = f(\varphi, \omega)$ and $(\eta_1, \dots, \eta_k) = g(\varphi, \xi)$, reducing inference to estimating $\Theta = (\beta, \sigma, p, \varphi, \omega, \xi)$.

2.2 Bayesian inference and prior distribution

For the bounded parameters, we use either a uniform or a Dirichlet. For the unbounded parameters, namely β and σ, we use a Jeffrey's prior on their joint distribution $\pi(\beta, \sigma) = 1/\sigma^2$. This prior enables us to derive the posterior distribution of β as a Gaussian distribution with parameters μ_β and Σ_β and the posterior distribution of σ as an inverse gamma distribution with parameters $n/2$ and $(\beta - \mu_\beta)^\top \Sigma_\beta^{-1}(\beta - \mu_\beta)/2$ with

$$\Sigma_\beta^{-1} = \frac{1}{\sigma^2} \sum_{j=1}^{k} \frac{1}{\tau_j^2} X_j^\top X_j, \qquad \mu_\beta^\top = -\frac{1}{\sigma^2} \left(\sum_{j=1}^{k} \frac{1}{\tau_j^2} (\sigma \alpha_j \mathbf{1}_{|I_j|} - Y_j)^\top X_j \right) \Gamma.$$

Using this model, we propose a robust Metropolis-within-Gibbs algorithm to estimate the model parameters. Our algorithm is adaptive and achieves a good mixing rate while ensuring convergence to the posterior distribution of the model parameters.

References

Alangari, N., Menai, M. E. B., Mathkour, H., and Almosallam, I. (2023). Intrinsically Interpretable Gaussian Mixture Model. *Information*, 14(3):164.

Armenian, S. H., Armstrong, G. T., Aune, G., Chow, E. J., Ehrhardt, M. J., Ky, B., Moslehi, J., Mulrooney, D. A., Nathan, P. C., Ryan, T. D., Van Der Pal, H. J., Van Dalen, E. C., and Kremer, L. C. (2018). Cardiovascular Disease in Survivors of Childhood Cancer: Insights Into Epidemiology, Pathophysiology, and Prevention. *Journal of Clinical Oncology*, 36(21):2135–2144.

Chen, Y., Georgiou, T. T., and Tannenbaum, A. (2019). Optimal Transport for Gaussian Mixture Models. *IEEE Access*, 7:6269–6278.

Chounta, S., Lemler, S., Haddy, N., Fresneau, B., Mansouri, I., Bentriou, M., Demoor-Goldschmidt, C., Diallo, I., Souchard, V., Do, T.-D., Veres, C., Surun, A., Doz, F., Llanas, D., Vu-Bezin, G., Rubino, C., De Vathaire, F., Letort, V., and Allodji, R. S. (2023). The risk of valvular heart disease in the French Childhood Cancer Survivors' Study: Contribution of dose-volume histogram parameters. *Radiotherapy and Oncology*, 180:109479.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34(2):187–202.

Erdmann, F., Frederiksen, L. E., Bonaventure, A., Mader, L., Hasle, H., Robison, L. L., and Winther, J. F. (2021). Childhood cancer: Survival, treatment modalities, late effects and improvements over time. *Cancer Epidemiology*, 71:101733.

Farewell, V. T. (1982). The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors. *Biometrics*, 38(4):1041.

Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors (2019). *Handbook of mixture analysis*. CRC Press, Taylor & Francis Group, Boca Raton.

Gudicha, D. W. and Vermunt, J. K. (2013). Mixture Model Clustering with Covariates Using Adjusted Three-Step Approaches. In Lausen, B., Van Den Poel, D., and Ultsch, A., editors, *Algorithms from and for Nature and Life*, pages 87–94. Springer International Publishing, Cham. Series Title: Studies in Classification, Data Analysis, and Knowledge Organization.

Hoshikawa, T. (2013). Mixture regression for observational data, with application to functional regression models. Version Number: 1.

Kamary, K., Lee, J. E., and Robert, C. P. (2018). Weakly Informative Reparameterizations for Location-Scale Mixtures. *Journal of Computational and Graphical Statistics*, 27(4):836–848.

Kunkel, D. and Peruggia, M. (2020). Anchored Bayesian Gaussian mixture models. *Electronic Journal of Statistics*, 14(2).

Maller, R., Resnick, S., Shemehsavar, S., and Zhao, M. (2024). Mixture cure model methodology in survival analysis: Some recent results for the one-sample case. *Statistics Surveys*, 18(none).

Murphy, K. and Murphy, T. B. (2020). Gaussian parsimonious clustering models with covariates and a noise component. *Advances in Data Analysis and Classification*, 14(2):293–325.

Papastamoulis, P. and Perrakis, K. (2024). Bayesian Cluster Weighted Gaussian Models. Version Number: 1.

Plate, T. A. (1999). Accuracy Versus Interpretability in Flexible Modeling: Implementing a Tradeoff Using Gaussian Process Models. *Behaviormetrika*, 26(1):29–50.

Safinianaini, N. and Boström, H. (2021). Towards interpretability of Mixtures of Hidden Markov Models. *arXiv*. Publisher: arXiv Version Number: 1.

Van Ness, M., Bosschieter, T., Din, N., Ambrosy, A., Sandhu, A., and Udell, M. (2023). Interpretable Survival Analysis for Heart Failure Risk Prediction. In Hegselmann, S., Parziale, A., Shanmugam, D., Tang, S., Asiedu, M. N., Chang, S., Hartvigsen, T., and Singh, H., editors, *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 574–593. PMLR.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15):1871–1879.

Zhang, Y., Wittrup, E., and Najarian, K. (2024). CoxFNN: Interpretable machine learning method for survival analysis. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5, Orlando, FL, USA. IEEE.

Synthèse de données par la méthode Avatar : fidélité en pharmacogénétique

C. Benoist[1], P. Marquet[1,2], J.-B. Woillard[1,2]

[1] *Service de pharmacologie, Toxicologie et pharmacovigilance, CHU de Limoges, France*

[2] *Pharmacologie & Transplantation, INSERM 1248, Université de Limoges, Limoges, France*

Auteur correspondant : Clément Benoist, Clement.BENOIST@chu-limoges.fr

Résumé. Pour que la ré-utilisation, à des fins de recherche, de données médicales soit conforme au RGPD (règlement général de la protection des données), il est nécessaire qu'elles soient totalement anonyminisées. La simple suppression des identités ne suffit pas à rendre les données anonymes. Notre idée est de générer des données à la fois anonymes et fidèles aux données originales pour partager les données. La possibilité d'évaluer et d'adapter le niveau de confidentialité au caractère sensible des données et à leur contexte d'utilisation conduit la CNIL (commission nationale de l'information et des libertés) à considérer l'algorithme Avatar comme conforme pour l'anonymisation des données au sens du RGPD. La fidélité sera considérée comme la capacité de l'algorithme à reproduire des résultats statistiques obtenus avec les données originales. Nous avons considéré une analyse de Cox en pharmacogénétique. Nous avons considéré différentes paramétrisations d'Avatar. Nous avons comparé les résultats avec l'algorithme classique CT-GAN (conditional tabular - generative adversarial network). Nous avons utilisé deux types de génération : sans et avec data augmentation. Avatar avec certains paramètres a des performances similaires à CT-GAN, mais permet de mieux respecter la confidentialité.

Mots-clés. Anonymisation, confidentialité, Avatar, données génératives, médecine

Abstract. So that medical data are not subject to GDPR (General Data Protection Regulation), it is sufficient that the data are anominized. The simple removal of identities does not necessarily render the data anonymized. Our idea is to generate data that is both anonymous and similar to the original, enabling safe sharing. The possibility to assess and adapt the privacy level to data sensitivity and their context of use led the French data protection authority (CNIL) to consider the Avatar method as compliant with anonymization in the sense of GPDR. The similarity is defined as the capacity of the synthetic data to give similar statistical results. We considered Cox survival analysis in pharmacogenetics. We considered different parametrizations of Avatar. We compared these results with CT-GAN (conditional tabular - generative adversarial network). There are two kinds of generation: with and without data augmentation. Well-parametrized Avatar has similar performance to CT-GAN, Avatar is privacy-focused.

Keywords. Anonymisation, privacy, Avatar, generative data, medicine.

L'article reprend une partie des résultats de [Woillard *et al.*, 2024]

1 Introduction

La confidentialité des données médicales est principalement régie par le règlement général de la protection des données (RGPD). En effet, la divulgation de données médicales (diagnostics, traitements,

séjours d'hospitalisation) peut avoir des conséquences majeures : discriminations à l'embauche, augmentation des primes d'assurances, stigmatisation sociale (VIH, maladies psychiatriques...)... Lorsque l'identité du patient n'est pas nécessaire à la recherche, il est possible de la remplacer par des pseudonymes. Néanmoins, cette méthode, bien qu'utile, n'est pas infaillible : il est possible de ré-identifier le patient. Même l'absence de pseudonymes formels ne garantit l'impossibilité de ré-identifier les patients car la combinaison de certaines variables peuvent le permettre. En ce qui concerne les données de santé, cette difficulté a été soulevée par Culnane *et al.* [2017] pour les bases de données de la sécurité sociale australienne rendues publiques, ce qui a entraîné sa mise en *offline* en urgence.

Une donnée non nominative sera considérée comme anonyme (ou anonymisée) quand elle ne permet pas de ré-identifier les patients, y compris par croisement avec d'autres bases de données publiques ou non. Dans le cas contraire, on parlera de données pseudonymisées (y compris lorsqu'il n'y a pas de pseudonymes formels dans la base de données.) Lorsque les données sont réellement anonymes, il est possible de les partager et de les publier librement ; notamment, ces données anonymes ne sont pas soumises au réglement général de la protection des données (RGPD). Ainsi, une méthode qui permettrait d'échanger des données nécessiterait donc de les anonymiser. On souhaite également que les données ressemblent aux données initiales (pseudonymisées), c'est-à-dire des données fidèles. Ainsi, l'idéal serait d'avoir des données à la fois anonymes et fidèles.

Une idée pour les générer des données fidèles serait d'utiliser des méthodes génératives. Les méthodes classiquement utilisées sont souvent issues du deep learning dont l'auto-encodeur variationnel (VAE) et le conditional tabular generative adversarial network (CT-GAN). Néanmoins, ces méthodes ne sont pas orientées confidentialité et souvent mal adaptées aux petits jeux de données. Une méthode générative validée par la Commission de l'informatique et des libertés (certification CNIL) est la méthode Avatar [Guillaudeux *et al.*, 2023] avec des conditions.

On peut définir la fidélité au sens fort comme le fait que les distributions statistiques des données initiales et des données synthétiques soient les mêmes. Néanmoins dans les statistiques appliquées, par exemple, à la médecine, il suffit que les résultats des analyses soient les mêmes (ou, du moins, proches des résultats initiaux). L'objectif de l'article est de vérifier que des résultats de pharmacogénétique (pharmacologie utilisant la génétique) soient les mêmes avec des données réelles et des données synthétiques.

Un exemple de problème de pharmacogénétique est de considérer les haplotypes du gène ABCB1 sur la perte du greffon rénal (rein d'un donneur reçu pour suppléer une défaillance rénale). En effet, lorsque le patient reçoit une greffe, se déclenche une réaction immunitaire contre le greffon, réaction qui est limitée par les médicaments immunosuppresseurs (anti-rejet). Il est montré que des mutations du gène ABCB1 du donneur peuvent favoriser une perte du greffon. Cet ensemble de mutations sur des bases nucléotidiques qui favorise la perte du greffon est 1236T/2677T/3435T, appelé de manière synthétique TTT. Cet ensemble de mutations existe en deux exemplaires (allèles). Le fait d'avoir deux allèles avec les mutations TTT favorise plus la perte du greffon rénal que le fait d'en avoir une seule. Le fait d'avoir un seule allèle avec les mutations TTT favorise plus la perte du greffon que l'absence totale de TTT pour les deux allèles.

2 Matériels et méthodes

2.1 Données

La base de données analysée provient d'une étude de pharmacogénétique qui lie le polymorphisme génétique de 227 donneurs avec les variables cliniques de leurs 253 receveurs du greffon rénal, receveurs traités par ciclosporine. Des modèles de Cox ont été développés pour évaluer l'association entre le polymorphisme génétique ABCB1 (mutation TTT) et le risque de perte du greffon. Rappelons que le modèle de Cox lie le risque instantané α de l'événement à l'instant avec les covariables $X_1, \ldots, X_n$ par la formule :

$$\alpha(t, X_1, \ldots, X_n) = \alpha_0(t) \exp\left(\sum_{i=1}^{n} a_i X_i\right),$$

où $\alpha_0(t)$ est le risque de base inconnu et généralement non estimé et les a_i sont des réels fixés inconnus à estimer. L'étude [Woillard *et al.*, 2010] montre l'association de la perte du greffon et, d'une part, la présence de deux haplotypes TTT (TTT/TTT vs autres haplotypes : hazard ratio (HR) = 9.346 (intervalle de confiance (IC) 95% : [2.278-38.461]), $p = 0.0019$) ; 1TTT vs 0TTT : 3.32, (IC95% : [0.92, 12.05]), p=0.0671) et, d'autre part, avec les antécédents de rejet aigu (HR=3.077 ; IC à 95%=[1.213,7.812] ; $p = 0.0178$). Les calculs ont été effectués en analyse multivariée (en considérant l'haplotype comme une variable ordinale). Dans notre étude, nous l'avons considéré comme une variable continue (0, 1, 2). Les données étudiées sont présentées de manière tabulaire avec une ligne par patient, il y a des variables continues et catégorielles. Il n'y a aucune valeur manquante dans les variables.

2.2 Avatar

L'algorithme considéré est l'algorithme Avatar. Les entrées de l'algorithme sont des données tabulaires pseudonymisées. Une analyse en composantes principales (ACP) permet de projeter ces données dans un espace latent. Dans l'espace latent (on ne conserve que les n_d premières composantes), on considère les k plus proches voisins. La donnée synthétique est une moyenne pondérée stochastique à partir de ces k plus proches voisins. La méthode peut synthétiser un nombre arbitraire de données synthétiques.

Soit O un individu dans l'espace latent. Pour ses k plus proches voisins, pour $i \in [\![1, k]\!]$, on considère D_i l'inverse de la distance entre O et le i-ème plus proche voisin. Soient $R_1, \ldots, R_n$ des réalisations de la loi exponentielle d'intensité λ. Enfin, $(C_i)_{i=1}^{k}$ est une permutation choisie aléatoirement de manière uniforme du k-uplet $(1/2, (1/2)^2, \ldots, (1/2)^k)$.

On calculera le poids brut P_i sous la forme du produit :

$$P_i = D_i \times R_i \times C_i.$$

Enfin, on normalise les poids bruts pour obtenir le poids final :

$$W_i = \frac{P_i}{\sum_{j=1}^{k} P_j}.$$

La donnée synthétique est calculée par combinaison linéaire associée aux poids W_i.

On répète cela pour toutes les données d'entrée. Éventuellement, le jeu de données d'entrée peut être parcouru plusieurs fois pour effectuer de la data augmentation.

Dans notre étude, nous considérerons des simulations avec les avatars avec des valeurs k du nombre de plus proches voisins de 5, 10 et 20. Pour chacune de ces valeurs, nous générons des données synthétiques, d'une part, de la taille des données originales et, d'autre part, nous générons des données synthétiques de quatre fois la taille des données synthétiques.

2.3 Conditional tabular-generative adversarial network (CT-GAN)

Les réseaux antagonistes génératifs s'appuient sur la mise en compétition de deux réseaux de neurones : d'une part, le générateur aura pour but de générer des données les plus ressemblantes possible aux données ; d'autre part, le discriminateur aura pour rôle de différencier les données synthétisées par le générateur et les données réelles.

Le CT-GAN s'appuie sur la formule des probabilités totales pour générer des données mixtes (incluant des variables catégorielles et les variables continues).

2.4 Étude de la variabilité à l'intérieur d'un jeu de données, entre les jeux de données et effet de la data augmentation

Nous avons évalué la variabilité d'Avatar à l'intérieur du jeu de données grâce à la méthode du bootstrap (100 réplications) du jeu de données synthétiques. En appliquant le modèle de Cox à chaque réplication, nous calculerons les quantiles 0.05, 0.5 et 0.95 du hazard ratio estimé relatif à la présence de l'haplotype TTT. Cette mini-étude sera produite pour les nombres de plus proches voisins précités.

De plus, nous évaluons la variabilité entre les jeux de données de l'algorithme Avatar. En effectuant plusieurs réalisations de la sortie de l'algorithme d'Avatar, en appliquant le modèle de Cox à chaque jeu de données, nous calculerons les quantiles 0.05, 0.5 et 0.95 du hazard ratio estimé relatif à la présence de l'haplotype TTT.

Nous investiguons l'effet de la data augmentation pour Avatar avec des nombres de plus proches voisins égaux à 5, 10 et 20 et nous comparons les résultats de l'analyse de Cox précitée avec variabilités à l'intérieur d'un jeu de données et entre les jeux de données. Les résultats statistiques du jeu de données augmenté (quatre fois la taille du jeu de données initial) sont comparées à celles du jeu de données synthétiques de la taille du jeu de données initial.

Les analyses effectuées avec les données synthétiques d'Avatar sont aussi effectuées avec CT-GAN, pour comparer les résultats à une méthode de référence.

3 Résultats

La table 1 présente le hazard ratio du modèle de Cox relativement à la présence de l'haplotype TTT. Chaque échantillon généré par une méthode générative a été répliqué 100 fois. Les quantiles

	Réalisation unique	Statistiques du HR après 100 bootstraps		
Méthode	HR (IC 95%)	q0.05	médiane	q0.95
Original	3.22 [1.70-6.09]	2.15	3.40	7.06
k=5	10.53 [4.47-24.82]	7.36	11.09	20.47
Augm. k=5	6.45 [4.38-9.50]	5.29	6.68	8.68
k=10	5.68 [2.61-12.36]	3.83	5.92	14.68
Augm. k=10	4.09 [2.68-6.20]	3.58	4.60	6.49
k=20	4.44 [1.5-13.14]	2.63	4.49	19.49
Augm. k=20	3.34 [1.97-5.67]	2.51	3.37	4.82
CT-GAN	3.20 [1.73-5.93]	2.00	3.50	7.77
Augm. CT-GAN	2.42 [1.91-3.08]	1.98	2.40	2.91

TABLE 1 : Statistiques du hazard ratio en fonction des méthodes de génération de données synthétiques. La deuxième colonne indique l'hazard ratio pour un seul jeu de données non « bootstrapé » avec l'intervalle de confiance à 95%. Les trois dernières colonnes présentent les statistiques (quantiles à 0.05 et à 0.95 et médiane) des HR pour 100 réplications du jeu de données réalisées par bootstrap. Les calculs sont effectués en analyse multivariée.

Algorithm	q0.05(HR)	médiane (HR)	q0.95(HR)
Original	.	3.40	.
k=5	4.07	6.17	9.35
Augm. k=5	4.44	5.53	7.27
k=10	3.98	6.43	12.84
Augm. k=10	4.24	5.30	7.76
k=20	3.49	5.98	12.90
Augm. k=20	3.58	5.30	7.25
CT-GAN	1.16	2.65	5.69
Augm. CT-GAN	1.46	2.50	4.45

TABLE 2 : Résultats de la variabilité des HR entre les jeux de données (100 générations indépendantes). Nous étudions la variabilité entre les jeux de données. Les calculs sont effectués en analyse multivariée.

à 0.05, 0.5 et 0.95 de ce hazard ratio ont été calculés.

La table 2 présente les statistiques (quantiles à 0.05, 0.5 et 0.95) du hazard ratio du modèle de Cox pour 100 générations indépendantes de chaque algorithme de synthèse de données. Dans les deux tables, les calculs sont effectués en analyse de Cox multivariée ; il y a d'autres variables explicatives que l'haplotype TTT.

4 Discussion

L'objectif de l'étude est d'évaluer la fidélité d'Avatar par rapport aux autres méthodes génératives. Nous nous plaçons dans un cas difficile où les données sont peu volumineuses ce qui est fréquent dans les études de santé et ce qui met en difficulté les méthodes de synthèse de données. La fidélité se définit comme la similarité des résultats statistiques avec ceux des données originales. Nous avons

une préférence nette pour Avatar du fait de sa certification CNIL, ce qui simplifie les démarches administratives.

Les tables 1 et 2 montrent, comme les données originales, la significativité de l'association entre l'haplotype TTT du donneur et la perte du greffon. L'association entre les deux variables est augmentée lorsqu'on utilise Avatar et diminuée lorsqu'on utilise CT-GAN. De plus, l'utilisation d'un jeu de données synthétiques avec data augmentation en réalisation unique, en réplication intra-jeu de données et en génération de plusieurs jeux de données (sauf, dans ce dernier cas, pour CT-GAN) a tendance à réduire le HR (ou sa médiane), ce qui aboutit à les résultats plus proches de l'original pour Avatar. De plus, comme attendu, la data augmentation réduit à la fois l'intervalle de confiance du hazard ratio (réalisation unique), sa variabilité intra-jeu de données et sa variabilité entre jeux de données générées indépendamment.

Augmenter le nombre de plus proches voisins k dans Avatar aboutit à des hazard ratio plus proches de l'hazard ratio original. En conclusion, le meilleur paramétrage pour Avatar est celui avec data augmentation et où $k = 20$. Néanmoins, les études plus poussées [Woillard *et al.*, 2024] font émerger le risque de faux positifs dans la sélection de variables par le modèle de Cox pour la data augmentation. Il fait aussi bien que CT-GAN avec une confidentialité suffisante selon le CNIL.

Remerciements

Ce travail fait partie du projet DIGPHAT, qui est soutenu par le gouvernement français, financé par l'agence nationale de la recherche (ANR) dans le cadre du programme France 2030 (référence : ANR-22-PESN-0017).

Nous remercions Julie Josse pour sa contribution.

Références

Culnane, C., Rubinstein, B. I., & Teague, V. (2017). Health data in an open world. *arXiv preprint arXiv :1712.05627*.

Guillaudeux, M., Rousseau, O., Petot, J., Bennis, Z., Dein, C. A., Goronflot, T., ... & Gourraud, P. A. (2023). Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *NPJ Digital Medicine*, 6(1), 37.

Woillard, J. B., Rerolle, J. P., Picard, N., Rousseau, A., Guillaudeau, A., Munteanu, E., ... & Marquet, P. (2010). Donor P-gp polymorphisms strongly influence renal function and graft loss in a cohort of renal transplant recipients on cyclosporine therapy in a long-term follow-up. *Clinical Pharmacology & Therapeutics*, 88(1), 95-100.

Woillard, J. B., Benoist, C., Destere, A., Labriffe, M., Marchello, G., Josse, J., & Marquet, P. (2024). To be or not to be, when synthetic data meet clinical pharmacology : A focused study on pharmacogenetics. *CPT : Pharmacometrics & Systems Pharmacology*.

SurvNMF: Non-Negative Matrix Factorization supervised for Survival Data Analysis

Vincent Le Goff [1,*], Vincent Guillemot [2], Cathy Philippe [3], Jean-François Deleuze [1], Edith Le Floch [1] & Arnaud Gloaguen [1]

[1] *Centre National de Recherche en Génomique Humaine (CNRGH), CEA, Université Paris-Saclay, Evry-Courcouronnes*
[2] *Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris*
[3] *Neurospin, CEA, Université Paris-Saclay, Gif-sur-Yvette*
[*] *vincent.legoff@cnrgh.fr*

Résumé. La factorisation matricielle non-négative est une méthode populaire d'analyse de données utilisée dans de nombreux domaines. En particulier, en biologie computationnelle, des extensions de la NMF utilisées pour estimer un espace réduit commun à un ensemble de matrices ont montré de bonnes performances pour retrouver des groupes d'individus. D'autre part, pour d'autres méthodes de réduction de dimension jointe, nous avons précédemment observé qu'une supervision de l'estimation de l'espace réduit peut améliorer la performance de ces méthodes pour la prédiction de la survie de patients atteints de cancers. Nous avons donc développé SurvNMF, un algorithme de NMF supervisé par la survie, capable d'apprendre de manière jointe une factorisation non-négative ainsi qu'un modèle de Cox sur l'espace latent engendré par la NMF. Des simulations montrent que SurvNMF peut améliorer la prédiction de survie dans certains cas, comparé à une NMF standard.

Mots-clés. NMF, analyse de survie, modèle de Cox, modèle joint

Abstract. Non-negative Matrix Factorization (NMF) is a popular data analysis method in many fields. In computational biology in particular, when NMF-based methods are used to estimate a joint reduced space out of a collection of matrices, they perform well at retrieving clusters. Furthermore, for other joint dimension reduction methods, we have previously observed that supervising the estimation of the reduced space can enhance the performance of such method when predicting survival of cancer patients. For these reasons, we developed SurvNMF, an algorithm that jointly learns a NMF and a Cox model on the latent space generated by the NMF. Through simulations, we have shown that SurvNMF outperforms unsupervised NMF in some situations.

Keywords. NMF, survival analysis, Cox model, joint modeling

1 Introduction

NMF is a popular matrix factorization technique that aims at decomposing a design matrix $\mathbf{X} \in \mathbb{R}_+^{n \times p}$ into the product of two non-negative matrices $\mathbf{W} \in \mathbb{R}_+^{n \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{p \times K}$: $\mathbf{X} \approx \mathbf{W}\mathbf{H}^\top$. This factorization is often found by minimizing the Frobenius norm of the residuals $\left\|\mathbf{X} - \mathbf{W}\mathbf{H}^\top\right\|_F^2$ (Lee et al., 2000).

Non-negative constraints in NMF lead to desirable clustering properties both in individual and variable space, which partially explains its popularity in a vast number of fields, especially in Computational Biology. Extensions of NMF to handle a collection of matrices have shown good performances in the very same field (Pierre-Jean et al., 2020). In the same multi-omics context, we benchmarked classical survival prediction methods with several joint Dimension Reduction (jDR) ones followed by a Cox regression model on the reduced space (Le Goff et al., 2024). Preliminary results show that jDR methods are particularly relevant in this context, especially when the reduction is supervised by survival data. Furthermore, among the unsupervised jDR methods, an extension of NMF performs particularly well.

Following that line, we developed SurvNMF, a supervised version of NMF for survival data that jointly learn the factorization with a Cox regression model on the matrix $\mathbf{W}$.

The Cox regression model is a semi-parametric model dedicated to explaining the link between covariates $\mathbf{x}_i \in \mathbb{R}^p$ and the time of an event $y_i \in \mathbb{R}_+$ corresponding either to failure ($\delta_i = 1$) or censoring ($\delta_i = 0$). It postulates that the instantaneous risk function for an individual, i.e. the probability that the event will appear for this individual at a given time knowing that it never happened before, can be written as $\lambda_0(t)\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)$, where $\lambda_0 : \mathbb{R}_+ \to \mathbb{R}_+$ is the base hazard. This separates the effect of time from the specificities of each individual. The regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ are determined by maximizing the Cox partial likelihood $\mathcal{L}$, that does not depend on λ_0

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i \in D} \frac{\exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right)}{\sum_{j \in R_i} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_j\right)}$$

with D the set of non-censored individuals and R_i the risk set of individual i, i.e. the set of individuals at risk (alive) immediately before failure of individual i. The negative partial log-likelihood is referred to as $\ell(\boldsymbol{\beta})$.

2 Algorithm for SurvNMF

To supervise the NMF, a joint model is learned between the factorization and a Cox model on the latent space $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$, leading to the following joint Cox-NMF problem

$$\begin{aligned} \underset{\substack{\mathbf{w}_i \in \mathbb{R}^n, \mathbf{h}_i \in \mathbb{R}^p, \\ \boldsymbol{\beta} \in \mathbb{R}^K}}{\operatorname{argmin}} \quad & f\left(\mathbf{h}_1, \mathbf{w}_1, \cdots, \mathbf{h}_K, \mathbf{w}_K, \boldsymbol{\beta}\right) = \frac{c_1}{2} \left\| \mathbf{X} - \sum_{k=1}^{K} \mathbf{w}_k \mathbf{h}_k^\top \right\|_F^2 + c_2 \ell\left(\mathbf{w}_1, \cdots, \mathbf{w}_K, \boldsymbol{\beta}\right) \\ & \text{s.t.} \quad \mathbf{w}_k \geq \varepsilon_{\mathbf{w}}, \mathbf{h}_k \geq \varepsilon_{\mathbf{h}}, \|\boldsymbol{\beta}\|_2^2 \leq s_{\boldsymbol{\beta}}, \|\mathbf{w}_k\|_2^2 \leq s_{\mathbf{w}} \end{aligned} \tag{1}$$

where $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$, $(c_1, c_2) \in \mathbb{R}_+^2$ are used to weigh the importance of each part of the criterion, $(\varepsilon_{\mathbf{h}}, \varepsilon_{\mathbf{w}}) \in \mathbb{R}_+^2$ are small (typically 10^{-8}) and used to enforce non-negativity, and $(s_{\mathbf{w}}, s_{\boldsymbol{\beta}}) \in \mathbb{R}_+^2$ are norm constraints used to prevent ℓ from going to $-\infty$.

An algorithm based on the Generalized Block Coordinate Descent (GBCD) framework (de Leeuw, 1994) is proposed here to solve (1), where each parameter ($\mathbf{w}_i, \mathbf{h}_i$ and $\boldsymbol{\beta}$) is updated in turn while keeping all the others fixed. Such a framework allows to design a monotone

algorithm with simple update rules that is guaranteed to converge towards a stationary point, under additional conditions not discussed in this document. The Hierarchical Alternating Least Square (HALS) algorithm was proposed for standard NMF and relies on a similar strategy (Cichocki et al., 2008).

Update for $\mathbf{h}_k$ Because the Cox log-likelihood does not depend on $\mathbf{h}_k$, the optimization problem associated with the minimization of f along $\mathbf{h}_k$ while every other parameters are fixed simplifies to a classical HALS:

$$\mathbf{h}_k^{t+1} = \underset{\mathbf{h}_k \geq \varepsilon_{\mathbf{h}}}{\operatorname{argmin}} \frac{1}{2}\left\|\mathbf{X}^{(k)} - \mathbf{w}_k^t \mathbf{h}_k^\top\right\|_F^2 = \mathcal{P}_{\varepsilon_{\mathbf{h}}}\left(\frac{\mathbf{X}^{(k)^\top}\mathbf{w}_k^t}{\|\mathbf{w}_k^t\|_2^2}\right)$$

with $\mathbf{X}^{(k)} = \mathbf{X} - \sum_{j \neq k} \mathbf{w}_j \mathbf{h}_j^\top$ and $\mathcal{P}_\varepsilon(\mathbf{x}) = \max(\varepsilon, \mathbf{x})$ the orthogonal projection of $\mathbf{x}$ on $]\varepsilon, +\infty[^p$.

Update for $\mathbf{w}_k$ To find an update for $\mathbf{w}_k$, we define the following function

$$f_{\mathbf{w}_k}(\mathbf{w}_k) = \frac{c_1}{2}\left\|\mathbf{X}^{(k)} - \mathbf{w}_k \mathbf{h}_k^\top\right\|_F^2 + c_2 \ell(\mathbf{w}_k)$$

with $\ell(\mathbf{w}_k)$ being $\ell(\mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\beta})$ with every variable but $\mathbf{w}_k$ fixed. In the framework of GBCD, we look for an update that will decrease the value of f, but not necessarily the best one. The computation of such update is based here on the Majorize-Minimize (MM) principle that defines a surrogate at the current point $\mathbf{w}_k^t$, i.e. a function $\tilde{f}_{\mathbf{w}_k}(\mathbf{w}_k, \mathbf{w}_k^t)$ such that $\tilde{f}_{\mathbf{w}_k}(\mathbf{w}_k^t, \mathbf{w}_k^t) = f_{\mathbf{w}_k}(\mathbf{w}_k^t)$ and $\forall \mathbf{w}_k \in \mathbb{R}^n, \tilde{f}_{\mathbf{w}_k}(\mathbf{w}_k, \mathbf{w}_k^t) \geq f_{\mathbf{w}_k}(\mathbf{w}_k)$. Then f is non increasing under the update rule

$$\mathbf{w}_k^{t+1} = \underset{\mathbf{w}_k}{\operatorname{argmin}}\, \tilde{f}(\mathbf{w}_k, \mathbf{w}_k^t)$$

Because $f_{\mathbf{w}_k}(\mathbf{w}_k^t) = \tilde{f}_{\mathbf{w}_k}(\mathbf{w}_k^t, \mathbf{w}_k^t) \geq \tilde{f}_{\mathbf{w}_k}(\mathbf{w}_k^{t+1}, \mathbf{w}_k^t) \geq f_{\mathbf{w}_k}(\mathbf{w}_k^{t+1})$.

Here, finding a quadradic majorizer of $f_{\mathbf{w}_k}$ only requires us to majorize $\ell(\mathbf{w}_k)$ because the NMF part of $f_{\mathbf{w}_k}$ is already quadratic. From de Leeuw (2016), such a majorizer exists if and only if the Hessian matrix $\nabla^2_{\mathbf{w}_k}\ell(\mathbf{w}_k)$ is uniformly bounded and therefore can be written as

$$\tilde{\ell}(\mathbf{w}_k, \mathbf{w}_k^t) = \ell(\mathbf{w}_k^t) + \left(\mathbf{w}_k - \mathbf{w}_k^t\right)^\top \nabla_{\mathbf{w}_k}\ell(\mathbf{w}_k^t) + \frac{1}{2}\left(\mathbf{w}_k - \mathbf{w}_k^t\right)^\top \mathbf{K}\left(\mathbf{w}_k - \mathbf{w}_k^t\right)$$

with $\mathbf{K} \in \mathbb{R}^{n \times n}$ chosen such that $\mathbf{K} - \nabla^2_{\mathbf{w}_k}\ell(\mathbf{w}_k)$ is positive definite for all $\mathbf{w}_k$. With $\|\cdot\|_{2,2}$ denoting the spectral norm, i.e. the greatest eigenvalue, and given the constraints on the original problem (1), it can be shown that $\nabla^2_{\mathbf{w}_k}\ell(\mathbf{w}_k)$ is indeed uniformly bounded as

$$\sup_{\mathbf{w}_k \in \mathbb{R}^n} \left\|\nabla^2_{\mathbf{w}_k}\ell(\mathbf{w}_k)\right\|_{2,2} \leq \exp\left(4K s_{\mathbf{w}} \|\boldsymbol{\beta}\|_2\right) \sum_{i \neq j} \sum_{m \in C_{ij}} \frac{\delta_m}{|R_m|^2} = u_{\mathbf{w}}$$

with C_{ij} the set of individuals that belong to both R_i and R_j. Finally, setting $\mathbf{K} = u_{\mathbf{w}}\mathbf{I}_n$ allows the condition of positive definiteness to be verified and we derive the update as

$$\mathbf{w}_k^{t+1} = \underset{\mathbf{w}_k \geq \varepsilon_{\mathbf{w}}, \|\mathbf{w}_k\|_2^2 \leq s_{\mathbf{w}}}{\operatorname{argmin}} \tilde{f}(\mathbf{w}_k, \mathbf{w}_k^t) = \mathcal{P}_{\varepsilon_{\mathbf{w}}}\left(\frac{c_1 \mathbf{X}^{(k)}\mathbf{h}_k^t + c_2\left(u_{\mathbf{w}}\mathbf{w}_k^t - \nabla_{\mathbf{w}_k}\ell(\mathbf{w}_k^t)\right)}{c_1\left\|\mathbf{h}_k^t\right\|_2^2 + c_2 u_{\mathbf{w}} + \lambda_{\mathbf{w}}}\right)$$

with $\lambda_{\mathbf{w}} \in \mathbb{R}_+$ used to enforce the ℓ_2-norm constraint and determined through binary search.

Update for $\boldsymbol{\beta}$ Similar arguments are used to find a surrogate $\tilde{\ell}(\boldsymbol{\beta}, \boldsymbol{\beta}^t)$ associated to an upper bound $u_{\boldsymbol{\beta}}$ for the Hessian matrix of ℓ alongside $\boldsymbol{\beta}$, leading to the following update

$$\boldsymbol{\beta}^{t+1} = \underset{\|\boldsymbol{\beta}\|_2^2 \leq s_{\boldsymbol{\beta}}}{\operatorname{argmin}} \tilde{\ell}(\boldsymbol{\beta}, \boldsymbol{\beta}^t) = \frac{u_{\boldsymbol{\beta}}\boldsymbol{\beta}^t - \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^t)}{u_{\boldsymbol{\beta}} + \lambda_{\boldsymbol{\beta}}}$$

This time, a closed form is available for $\lambda_{\boldsymbol{\beta}} \in \mathbb{R}_+$ that enforces the norm constraint.

Algorithm 1: Algorithm for SurvNMF

Data: $\mathbf{X}, (y_1, \delta_1), \ldots, (y_n, \delta_n)$
Parameters: $K, c_1, c_2, \varepsilon_{\mathbf{w}}, \varepsilon_{\mathbf{h}}, s_{\mathbf{w}}, s_{\boldsymbol{\beta}}$
Result: $\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}$
Initialize: $\mathbf{W} \in \mathbb{R}^{n \times K}, \mathbf{H} \in \mathbb{R}^{p \times K}, \boldsymbol{\beta} \in \mathbb{R}^K$

repeat
  for $k \leftarrow 1$ **to** K **do**
    $\mathbf{X}^{(k)} \leftarrow \mathbf{X} - \sum_{j \neq k}^{K} \mathbf{w}_j \mathbf{h}_j^\top$;
    $\mathbf{h}_k \leftarrow \mathcal{P}_{\varepsilon_{\mathbf{h}}}\left(\frac{{\mathbf{X}^{(k)}}^\top \mathbf{w}_k}{\|\mathbf{w}_k\|_2^2}\right)$;
    $u_{\mathbf{w}} \leftarrow \exp\left(4Ks_{\mathbf{w}}\|\boldsymbol{\beta}\|_2\right) \sum_{i \neq j} \sum_{m \in C_{ij}} \frac{\delta_m}{|R_m|^2}$;
    $\mathbf{w}_k \leftarrow \mathcal{P}_{\varepsilon_{\mathbf{w}}}\left(\frac{c_1 \mathbf{X}^{(k)} \mathbf{h}_k + c_2\left(u_{\mathbf{w}}\mathbf{w}_k - \nabla_{\mathbf{w}_k}\ell(\mathbf{w}_k)\right)}{c_1\|\mathbf{h}_k\|_2^2 + c_2 u_{\mathbf{w}}}\right)$;
    if $\|\mathbf{w}_k\|_2^2 > s_{\mathbf{w}}$ **then**
      Determine $\lambda_{\mathbf{w}}$ with binary search;
      $\mathbf{w}_k \leftarrow \mathcal{P}_{\varepsilon_{\mathbf{w}}}\left(\frac{c_1 \mathbf{X}^{(k)} \mathbf{h}_k + c_2\left(u_{\mathbf{w}}\mathbf{w}_k - \nabla_{\mathbf{w}_k}\ell(\mathbf{w}_k)\right)}{c_1\|\mathbf{h}_k\|_2^2 + c_2 u_{\mathbf{w}} + \lambda_{\mathbf{w}}}\right)$;
    end
  end
  $u_{\boldsymbol{\beta}} \leftarrow \sum_{i \neq j} \sum_{m \in C_{ij}} \delta_m \frac{\exp(s_{\boldsymbol{\beta}}\|\mathbf{w}_{i\cdot}\|_2)\exp(s_{\boldsymbol{\beta}}\|\mathbf{w}_{j\cdot}\|_2)}{\left(\sum_{l \in R_m} \exp\left(-s_{\boldsymbol{\beta}}\|\mathbf{w}_{l\cdot}\|_2\right)\right)^2}$;
  $\lambda_{\boldsymbol{\beta}} \leftarrow \max\left(0, \frac{\left\|u_{\boldsymbol{\beta}}\boldsymbol{\beta} - \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})\right\|_2^2}{\sqrt{s_{\mathbf{w}}}} - u_{\boldsymbol{\beta}}\right)$;
  $\boldsymbol{\beta} \leftarrow \frac{u_{\boldsymbol{\beta}}\boldsymbol{\beta} - \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})}{u_{\boldsymbol{\beta}} + \lambda_{\boldsymbol{\beta}}}$;
until *difference of criterion between iterations is inferior to* 10^{-8};

The constants c_1 and c_2 are respectively set to $\frac{(1-\theta)}{\sqrt{Q}}$ and $\theta\sqrt{Q}$, where Q controls the balance between the two parts of optimization criterion (1). In practice, to estimate the value of Q, we run algorithm 1 with $Q = 1$ for 10 iterations. $\theta \in [0, 1[$ controls the balance between decomposition and supervision in the algorithm. In many cases, without this procedure, the algorithm will overfit the survival data while ignoring the NMF part of the criterion.

As usual in NMF, we run multiple random starts. For computational reasons, they were run for a fixed number of iterations with an approximated algorithm using a quasi-Newton update for $\mathbf{w}_k$, before picking the best random start and continuing with algorithm 1.

3 Simulation experiments

Data generation $\mathbf{W} \in \mathbb{R}^{n\times K}$ and $\mathbf{H} \in \mathbb{R}^{p\times K}$ are drawn from a standard normal distribution, an offset is added to make them non-negative and each column (resp. line) of $\mathbf{W}$ (resp. $\mathbf{H}$) is divided by its ℓ_2-norm. From the K components generated, only K^* contribute to survival. Defining $\mathbf{X}_{\text{signal}}$ the matrix generated by K^* survival components and $\mathbf{X}_{\text{noise}}$ the one generated by the $K - K^*$ non-survival components, we compute the full matrix $\mathbf{X} = \sqrt{\text{SNR}}\frac{\mathbf{X}_{\text{signal}}}{\|\mathbf{X}_{\text{signal}}\|_F^2} + \frac{\mathbf{X}_{\text{noise}}}{\|\mathbf{X}_{\text{noise}}\|_F^2}$, with SNR the desired Signal (survival) to Noise Ratio.

$\boldsymbol{\beta}$ is drawn from a standard normal distribution and values corresponding to components that do not participate to survival are set to 0. Finally, survival times are simulated with an exponential distribution and censoring is drawn from a Bernoulli distribution $\mathcal{B}(0.1)$.

We simulate $n = 375$ observations, with $p = 3000$ variables, and $K = 5$ components. We then consider all combinations of $K^* \in \{1, 3\}$ and SNR $\in \{0.005, 0.05, 0.5\}$. Each setting is replicated 20 times.

Considered algorithms For SurvNMF, all combinations of the number of components $\hat{K} \in \{1, 3\}$ and $\theta \in \{0, 0.1, 0.5, 0.9\}$ were tested. We also set $s_{\mathbf{w}} = s_{\boldsymbol{\beta}} = 1$. SurvNMF is compared to the NMF implemented in the R package NMF (Gaujoux et al., 2010) with algorithm 'lee' that uses a multiplicative update (latter referred to as NMF_Lee) and the same values of $\hat{K}$. 20 random starts were run for each algorithm and a Cox model was adjusted each time on $\mathbf{W}_{\text{train}}$, the latent space on the train data.

Models were systematically trained on 300 (80%) individuals and tested on 75 (20%). Estimation of the latent space on the test set $\mathbf{W}_{\text{test}}$ is done by minimizing a least square criterion for both SurvNMF and NMF_Lee while imposing the appropriate constraints through an HALS algorithm. We assess the performance of both algorithms to predict survival with Uno's C-index (Uno et al., 2011) using $\mathbf{W}_{\text{test}}$.

Results Results are shown in figure 1. First, when $\hat{K} = 1$, NMF_Lee and SurvNMF lead to similar results, independently of the values of θ or K^* and in general to poorer results than when $\hat{K} = 3$. Furthermore, C-index increases with the SNR. In the following, we detail the results regarding $\hat{K} = 3$.

For low SNR values, two atypical situations (NMF_Lee, $K^* = 1$ and $\theta = 0$, $K^* = 3$) need further investigation. They might be linked to a higher sensitivity to initialization in such low SNR situation. In general, low values of C-index are predicted. For intermediate SNR, survNMF outperforms NMF_Lee especially for $K^* = 3$. This improvement is much lower when $\theta = 0$ (unsupervised) is compared to other θ values. For high SNR, high supervision essentially has a negative impact and seems to lead to overfitting.

Perspectives Future developments will aim at selecting methods able to determine optimal values for θ and Q, for instance with cross-validation. Necessary conditions for our GBCD algorithm to converge towards a stationary point will be verified in the near future.

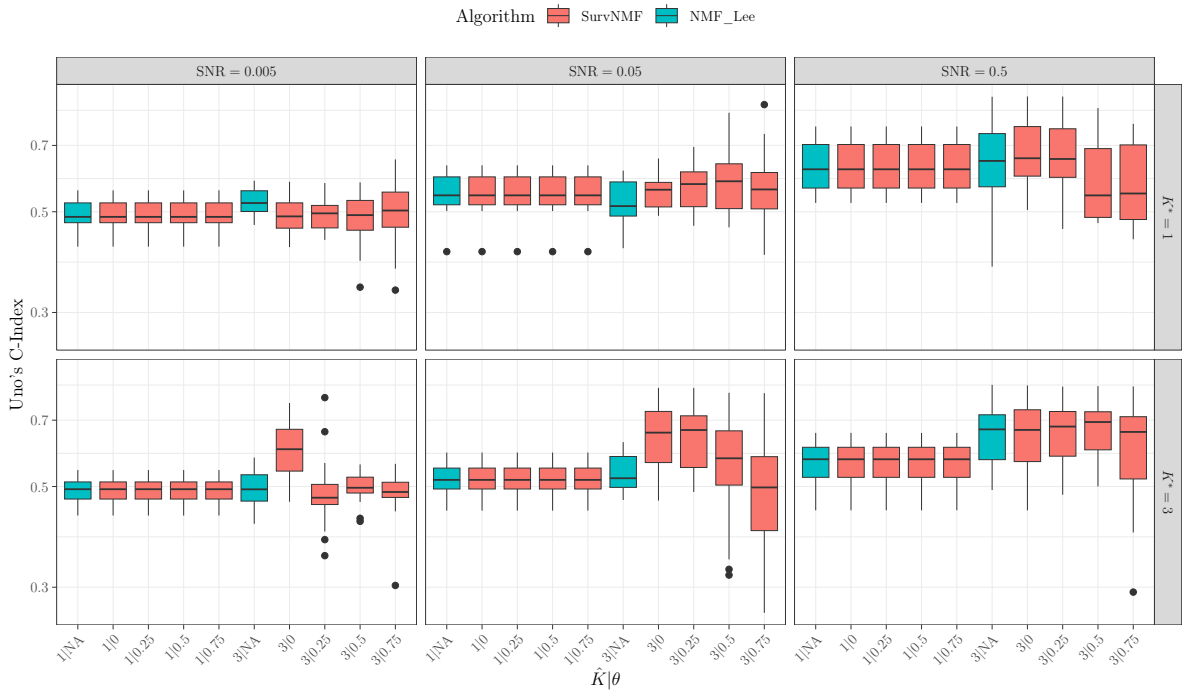


Figure 1: C-Index on the test set $\mathbf{W}_{\text{test}}$. Top to bottom: $K^* = 1$ and $K^* = 3$. Left to right: SNR = 0.005 (high noise), SNR = 0.05 (intermediate) and SNR = 0.5 (low).

References

Cichocki, A. et al. (2008). Flexible HALS algorithms for sparse non-negative matrix/tensor factorization. In *2008 IEEE Workshop on Machine Learning for Signal Processing*.

de Leeuw, J. (1994). Block-relaxation Algorithms in Statistics. In *Information Systems and Data Analysis*, pages 308–324, Berlin, Heidelberg. Springer.

de Leeuw, J. (2016). *Block Relaxation Methods in Statistics*. https://bookdown.org/jandeleeuw6/bras/.

Gaujoux, R. et al. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, 11(1):367.

Le Goff, V. et al. (2024). Impact of joint Dimension Reduction methods for survival prediction - Extension of a multi-omics benchmark study. https://cea.hal.science/cea-04734003.

Lee, D. et al. (2000). Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Pierre-Jean, M. et al. (2020). Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. *Briefings in Bioinformatics*, 21(6):2011–2030.

Uno, H. et al. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117.

Challenging estimation of tuning parameters in sPLSDR-Cox models to predict the incidence of rare diseases

Ian Meneghel Danilevicz[1*†] Sam Vidil[1*] Séverine Sabia[1,2*]

[1] Epidemiology of Ageing and Neurodegenerative Diseases, Université Paris-Cité, INSERM, U1153, CRESS, 10 Avenue de Verdun, Paris, 75010, France
[2] UCL Brain Sciences, Division of Psychiatry, University College London, London, UK

Keywords: Machine learning, Survival analysis, Variable selection, Cross-validation, Bootstrap.

Abstract: We investigate some practical points of the sparse partial least squares deviance residuals Cox regression (sPLSDR), a machine learning (ML) method with great application potential and a powerful alternative for Cox regression analysis when dimension reduction and variable selection are required. It can be used in a broad range of contexts, among which the health domain. This method requires to tune two hyper-parameters: eta (sparsity level) and D (number of components). Usually, the literature recommends cross-validation (CV) with a number of folds that depends on the sample size to estimate the tuning parameters. However, in case of dichotomous outcome for which the ratio of cases and non-cases is small, as in rare disease such as incidence of major neurocognitive disorder, this approach appears to be unstable. In that case, even large samples may perform as poorly as small samples. We specifically test this scenario of large samples with rare events using Monte Carlo simulations with different performance criteria (C-index and AIC). We observed that five folds are more conservative (biased) than one fold, and it presents more variability (less precise) in case of rare event outcomes. Previous authors suggested that Bootstrap or leave-one-out (LOO) could be a wise alternative. We checked Bootstrap with one fold in our real analysis application using the Whitehall II accelerometer sub-study (N total = 3,969, cases = 291 (7.3%)) as a training set and the UK Biobank as a validation set (N total = 54,833, cases = 750 (1.4%)), both considerably large sample size studies, with incidence of major neurocognitive disorders as a example of rare outcomes. This application shows that CV is sensitive to the number of folds and the initial seed. The counterpart Bootstrap was more robust and displayed two components with eta equal to 0.75 as a consistent optimum choice for the tuning parameters. However, this approach requires external validation, which is a limitation to this approach as it is not always feasible. Further studies can better map the ideal number of folds for CV depending on the sample size and the ratio of positive cases. As dichotomous outcomes and survival time are classical statistical subjects, the ML literature must provide clear and precise recommendations for several available dawn methods.

1 Background

Accelerometer devices objectively measure the wrist acceleration of human participants [19]. The derived features calculated from the accelerometer three axial time series aggregated per minute for each participant during a week can be used to estimate rest-activity rhythm, daytime activity, sleep, and chronotype. These features can be used to predict several health outcomes such as mortality [8], cardio-metabolic diseases [6], and neurodegenerative diseases [20]. This represents a vast potential for medical research [7] as these devices provide a more objective and detailed information than commonly used self-reported data.

In survival analysis, we model the duration from a time origin until an event of interest occurs, considering a set of predictors. In case of several candidate predictors, some selection and/or combination of predictors are often operated to allow for a parsimonious model, using machine learning (ML) methods such as the sparse partial least squares deviance residuals Cox (sPLSDR) [2].

ML methods often require tuning parameters to be fitted. These parameters are usually pre-estimated using cross-validation (CV) techniques. The general recommendation consists of doing 3 to 10 folds (depending on the sample size) [15], within a CV with all combinations of hyper-parameters. The sPLSDR requires two hyper-parameters: 1) $\eta \in [0, 1)$, where 0 means a full dense model with all features, and values close to 1 a highly sparse model with only few features retained; 2) D components (each component is a combination of features as in PLS regression). One approach consists in testing all values of $\eta = 0, 0.05, 0.10, \ldots, 0.95$, which is 20 values, combined with $D = 4$, returning 80 model fits for each fold. Then, a finer scale of eta (for example, $\eta = 0.50, 0.51, 0.52, \ldots, 0.59$) can be used after identifying a region of interest based on some criteria, eg, integrated AUC survival ROC (iAUC) [16]. This approach is commonly used, but there are some cases of prominent instability in this process.

*Contacts: ian.meneghel-danilevicz at inserm.fr, sam.vidil at inserm.fr, s.sabia at ucl.ac.uk
†Speaker

It happens that when the outcome of interest is rare, such as for example major neurocognitive disorder (NCD), the process of optimisation is unstable for different fold values, and even for the same fold value, it may vary according the initial seed, indicating a lack of reliability. This instability could be due to small sample size where 5 folds for CV are recommended, but we hypothesise that the low proportion of cases even in cases of large samples could also be problematic.

We plan to discuss ways to overcome the challenge of tuning parameter estimation in cases of rare events. Our application is based on the identification of accelerometer data predicting major neurocognitive disorder (NCD), such as Alzheimer's disease, using the Whitehall II ($n = 3969$) as a training set and the UK Biobank (UKB) as a validation set. Our proposition to get stable results for the tuning parameters is to keep all the data from the training set and bootstrap it, and then to test the validity of the model using in an external dataset (UKB). The optimization process for Whitehall II is unstable for different fold values, and even for the same fold value (with different initial seed), which indicates a lack of reliability. The instability is not due to a small sample size, as Whitehall II is a relatively large cohort. However, the outcome presents fewer cases than non-cases (the target disease is the major neurocognitive disorder (NCD), which includes Alzheimer's disease). The problem was not fixed even when considering the rules for a small sample size (5 folds).

A small simulation study was conducted to test the hypotheses of low stability in case of small ratio between cases and non-cases. We artificially created data sets with $n = 5000$ where the maximum observation time is 10 years of a disease whose incidence over time is small (resulting in few cases occurred during the observation period). In these simulations, we divided the data into five folds or keep the whole pool together, and we estimated the models using traditional performance criteria, such as the C-index and the Akaike information criterion (AIC).

2 Methods and data

2.1 The sPLS technique

The sparse partial least squares regression (sPLS) was proposed by [9] to promote variable selection during PLS dimension reduction. The sPLS can select variables for the first direction vector and additional variables in the construction of the subsequent direction vectors [2].

A direct extension of PLS regression to sPLS regression could be provided by imposing L_1 constraint on PLS direction vector $\boldsymbol{w}$ as

$$\max_{\boldsymbol{w}} \boldsymbol{w}^T \boldsymbol{M} \boldsymbol{w} \text{ subject to } \boldsymbol{w}^T \boldsymbol{w} = ||\boldsymbol{w}||_2 = 1, ||\boldsymbol{w}||_1 \leq l, \tag{1}$$

where $\boldsymbol{M} = \boldsymbol{X}^T \boldsymbol{Y} \boldsymbol{Y}^T \boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{X}$ are matrices with the same number of rows, $l > 0$ is the L_1 constraint.

2.2 The Cox proportional hazards model

The model assumes the following hazard function for the occurrence of an event at time t in the presence of censoring as follows $h(t) = h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{X})$, where $h_0(t)$ is a baseline hazard function, $\boldsymbol{\beta}$ is a vector of regression coefficients and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ is a $n \times p$ matrix of features with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ a row-vector for the i-th individual. The event could be any disease of interest. Based on the available data, Cox's partial likelihood can be written as

$$L(\boldsymbol{\beta}) = \prod_{k \in \Omega} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_k)}{\sum_{j \in \zeta_k} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_j)},$$

where Ω is the set of indices of the events and ζ_k denotes the set of indices of the individuals at risk at time t_k. The goal is to find the coefficients which maximize the log partial likelihood function [2].

2.3 Deviance residuals

Here, we consider a Cox model with no time-dependent explanatory variables and, at most, one event per patient. Regardless, the Martingale residuals for the i-th subject with observation time t_i and event status δ_i, where $\delta_i = 0$ if t_i is a censored time, and $\delta_i = 1$ otherwise is as $\hat{M}_i = \delta_i - \hat{\Delta}_0(t_i) \exp(\hat{\boldsymbol{\beta}}^T x_i)$, where $\hat{\Delta}_0(t_i)$ is the estimated cumulative hazard function at time t_i [2]. However, the Martingale residuals used to be highly skewed [2]. The deviance residuals d_i are a normalized transform of the martingale residuals. For the Cox model, the deviance residuals [10] amount to a form as

$$d_i = \text{sign}(\hat{M}_i)[2\{-\hat{M}_i - \delta_i \log(\delta_i - \hat{M}_i)\}]^{0.5}, \tag{2}$$

where sign is the signal.

2.4 The sPLSDR method

The sparse partial least squares deviance residual Cox (sPLSDR) method was proposed by [2] as an upgrade concerning sPLS proposed by [9] and partial least squares Cox (PLS-Cox) [3, 1]. The general idea is to apply the sPLS technique presented in 1 on the deviance residual from equation 2.

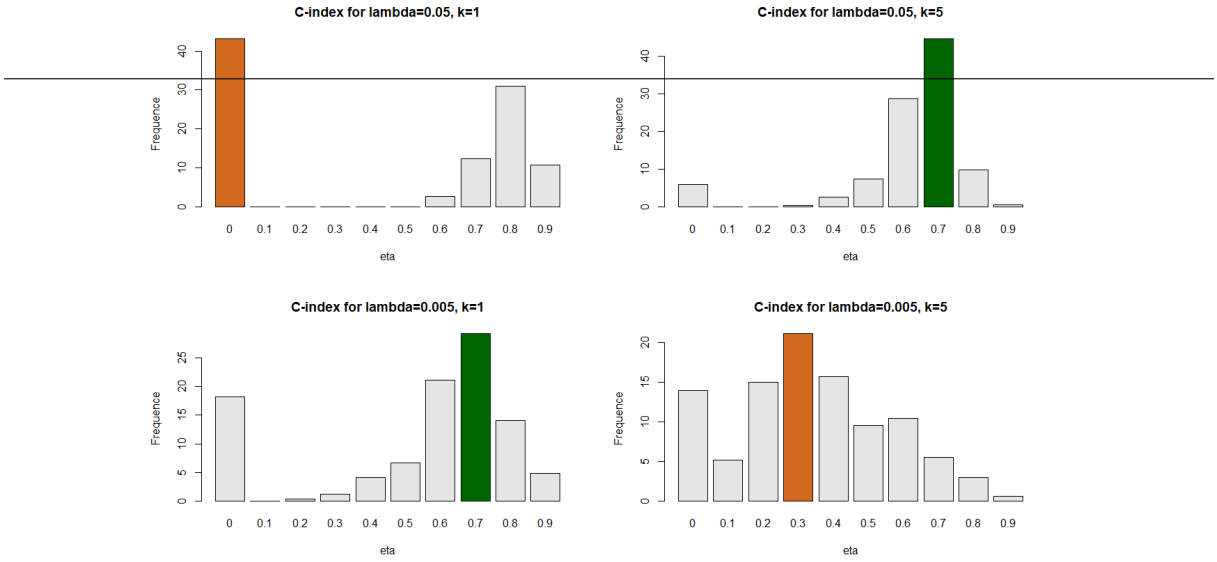


Figure 1: Frequency distribution of η as a function of C-index for $\lambda = 0.05$ on top and $\lambda = 0.005$ on bottom, for $k = 1$ on left and $k = 5$ on right. Colored bars correspond to the highest frequency, orange when $\eta < 0.5$, and green when $\eta > 0.5$.

2.5 Monte Carlo simulation

We used an exponential survival model, where the hazard function is $h_i(t) = \lambda \exp(\boldsymbol{X}_i^T \boldsymbol{\beta})$, and the cumulative hazard is $H_i(t) = \lambda t \exp(\boldsymbol{X}_i^T \boldsymbol{\beta})$, as implemented on the **simsurv** R Package by [5]. We create artificial data with $n = 5000$ number of individuals, $\lambda = 0.05, 0.005$ rate parameter, $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\beta}_3^T)$ is the coefficient parameters, composed by $\boldsymbol{\beta}_1^T = (0.15, 0.1, 0.05)$, $\boldsymbol{\beta}_2^T = (-0.15, -0.1, -0.05)$ and $\boldsymbol{\beta}_3^T = (0, 0, 0, 0, 0, 0)^T$; the covariate matrix is $\boldsymbol{X}_{i,(1:3)} \sim N(\boldsymbol{\mu}_1, \boldsymbol{R})$ a multivariate normal with $\boldsymbol{\mu}_1 = (0.15, 0.1, 0.05)^T$ and $\boldsymbol{R}$ a matrix with variance on the diagonal of ones and covariance on the off-diagonals of 0.5; $\boldsymbol{X}_{i,(4:6)} \sim N(\boldsymbol{\mu}_2, \boldsymbol{R})$ a multivariate normal with $\boldsymbol{\mu}_2 = (-0.15, -0.1, -0.05)^T$ and same covariance; $\boldsymbol{X}_{i,(7:12)} \sim N(\boldsymbol{\mu}_3, \boldsymbol{I})$ a multivariate normal with $\boldsymbol{\mu}_3 = (0, 0, 0, 0, 0, 0)^T$ and $\boldsymbol{I}$ an identity matrix. The maximum observed time is $m = 10$.

So this simple model may be well estimated by a sPLSDR model with two dimensions and a considerable sparsity ($\eta > 0.5$). We proceed with a Monte Carlo simulation, replicating $M = 1000$ times each λ and estimating the model changing $\eta = 0, 0.1, 0.2, \ldots, 0.9$, using $k = 1, 5$ folds. We save the C-index and the AIC of each iteration of the Monte Carlo. We set all models with two dimensions. Finally, we count how many times each value of η was considered the best in the row; this results in a frequency per value of η.

2.6 Real datasets

The Whitehall II study is an ongoing prospective cohort study established in 1985-1988 among 10,308 British civil servants with clinical examinations every four–five years since inception [18]. An accelerometer measure was added to the 2012–2013 wave of data collection, where 4,880 members from the cohort were invited to take part in the sub-study. Participants were followed for incident NCD up to 2023.

The UK Biobank (UKB) is a prospective population-based cohort study on over 500,000 men and women aged between 40 and 69 years established in 2006-2010 [17]. The accelerometer sub-study was proposed to 236,488 individuals between 2013-2015 [12]. Participants were followed for incident NCD up to 2022.

The accelerometer signal from each participant in both cohorts were extracted by **GGIR** R package [19]. A brief introduction to the 40 accelerometer derived features can be found in [8, 11]. They represent features related to rest-activity rhythm, daytime activity, sleep, and chronotype.

When there are two datasets, usually the bigger is used as a training set and the smaller as a validation set [15, page 81]. However, we intend to test our hypotheses in separate subpopulations, so we choose Whitehall II ($n = 3969$; 291 incident cases (7.3%)) as a training set in order to validate findings in UKB ($n = 54833$; 750 incident cases (1.4%)) and subgroups of this population. This should not be a problematic because Whitehall II is not a small cohort.

3 Results

3.1 Simulation results

Figure 1 shows the frequency decision taken by the highest C-index, ie at each simulation, we count the η, which corresponds to the highest C-index. On top, $\lambda = 0.05$, which corresponds to a scenario where there are several incident

Table 1: Median number of features per η in each simulation set with $\lambda = 0.05, 0.005$.

η	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\lambda = 0.005$	12	12	12	12	12	11	9	7	5	4
$\lambda = 0.05$	12	12	12	12	12	12	12	11	8	4

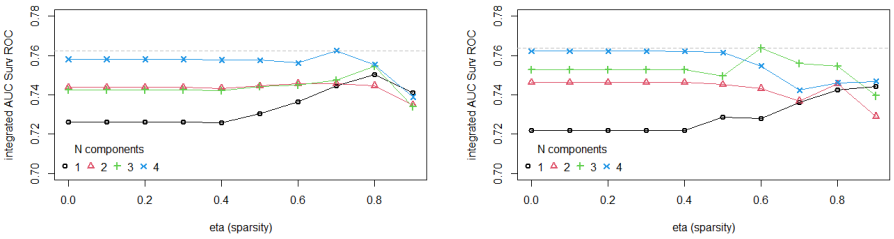


Figure 2: The iAUC for $\eta = 0, 0.1, \ldots, 0.9$ for n components and $k = 5$ (left and right panel differ by the initial seed used).

cases (eg, common disease), and on the bottom, $\lambda = 0.005$ is a scenario with few incident cases (eg, rare disease). On the left, the estimation is done by a $k = 1$ fold, and on the right, a $k = 5$ folds does it. For $\lambda = 0.05$ the most frequent decision is $\eta = 0.7$ when using $k = 5$ which is in line with the general recommendation and with the complexity of the model, but using $k = 1$ the most frequent decision is $\eta = 0$ (too conservative decision). However, when $\lambda = 0.005$ the decision is the reverse: using $k = 1$ returns $\eta = 0.7$ as the most frequent decision and $k = 1$ returns $\eta = 0.3$ as the most frequent decision (which is conservative). Furthermore, we can see that the plot at the top right ($k = 5$) presents larger frequency variability than $k = 1$. We also show in the Table 1 the corresponding median number of features retained as a function of η. For $\eta \leq 0.5$ all the models are highly dense, conserving almost all features in median, which is not in line with the simulation (as the vector $\boldsymbol{\mu}_3$ only has zeros, further $\boldsymbol{X}_{i,(1:3)}$ and $\boldsymbol{X}_{i,(4:6)}$ are inner correlated). Similar findings were observed when using AIC instead of C-index as the performance criteria (data not shown).

3.2 Real data results

Figure 2 shows the iAUC for $\eta = 0, 0.1, \ldots, 0.9$ for $D = 1, \ldots, 4$ (components) from 5-fold CV. At the left, the maximum is achieved by D=4 components at $\eta = 0.7$ and at the right by $D = 3$ at $\eta = 0.6$. The only difference is the seed to sampling the participants in each fold. This variability was observed irrespective of the number of folds used for CV from $k = 3$ to 10 folds.

Then in order to identify the best tuning parameters that allow the selected model to improve the predictive value for NCD as compared to known risk factors (sex, age, ethnicity, education, working status, cohabitation status, smoking, alcohol consumption, daily consumption of fruits and vegetables, BMI, hypertension, diabetes, hyperlipidemia, multiple comorbidities, central nervous system (CNS) medication), we compared the C-index of the model with only the known risk factors (dotted line) with the C-index of models including these risk factors and the fitted accelerometer data for $\eta = 0, 0.05, \ldots, 0.95$ for $D = 1, \ldots, 4$, Figure 3. The C-index mean (points) and CI (bars) were estimated using bootstrap with 1 to 4 components derived from the accelerometer features and all the previously described known risk factors. The C-index for the covariates only was around 0.78. We found that 2 components significantly higher than known risk factors only for any $\eta \in [0, 0.75]$. So the model retained was the one for $D = 2$ and $\eta = 0.75$, allowing to improve C-index while being as sparse as possible.

We created 2 composite acceleration scores in the UKB using features and their loadings identified in the Whitehall

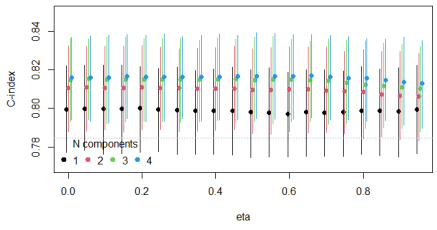


Figure 3: C-index (mean and CI) for $\eta = 0, 0.05, \ldots, 0.95$ for 1 to 4 accelerometer components, $k = 1$, with covariates using Bootstrap with 2000 repetitions (dotted line represents the C-index of socio and health covariates only).

II cohort. The first composite score was able to improve the predictive value of a model including known risk factors in this external cohort, but it was not the case for the second score.

4 Discussion

The Monte Carlo simulation shows that one-fold is more stable than five if the cases are rare. Indeed, $\lambda = 0.005$ associated with $k = 5$ display more conservative decisions ($\eta < 0.5$) and larger variability than $\lambda = 0.005$ associated with $k = 1$. In case of rare events, identifying a stable decision of hyper-parameters may be a challenge. Our findings suggest that it may be more reliable to use one fold, opening to two isuses: 1) if there is no repetition, we cannot estimate a confidence interval (CI); 2) we cannot validate our findings as there is no remaining data to test. This raises the importance of having an independent datasets as validation sample. The necessity of external validation is a limitation of this approach, as it is not always available.

Our real data application shows that CV is unstable for $k = 5$ in a large sample size as Whitehall II ($n = 3969$). Changing only the initial seed, CV for $k = 5$ returns contradictory results such as different optimum values for η and D. The bootstrap result shows a very stable shape for η. The conclusion was that $D = 2$ with $\eta = 0.75$ significantly improved (higher C-index) compared with a model including known risk factors only. Findings were in part replicated in an external validation (UKB), where the first component identified also improved the C-index of a model including known risk factors. Several reasons might explain the lack of significance of the second component : UKB participants are relatively younger and there is a lack of a sleep diary for UKB implying higher measurement errors in the measures of sleep that are particularly loaded in the second component scores.

Applied works commonly used 5 folds [8] or 10 folds [13, 14]. A few studies give general advice for CV for similar machine learning techniques; they can recommend between 2 and 10 folds [4] or 3 and 10 folds [15, page 82] depending on the sample size. However, they do not consider the ratio of cases by non-cases in dichotomous outcomes. Lê Cao and Welham suggest leave-one-out (LOO) for tiny samples and that Bootstrap might be an interesting new approach [15, page 82]. Our simulation and real application agree that $k = 1$ is a promising approach for specific situations with few positive outcomes (eg, rare diseases).

5 Conclusion

The general advice for CV proceedings takes into account the sample size. However, from our study, the number of cases (in dichotomous outcomes) also plays an important role and should not be neglected. Furthermore, dividing the data by folds to CV does not seem a stable strategy if the ratio of cases/noncases is too extreme (which is the case for rare diseases). In fact, a bootstrap on the entire sample appears to be a good solution. However, this approach requires external validation, which is not always available. Further studies can confirm other scenarios where this situation occurs.

References

[1] P. Bastien. Deviance residuals based PLS regression for censored data in high dimensional setting. *Chemometrics and Intelligent Laboratory Systems*, 91(1):78–86, 2008.

[2] P. Bastien, F. Bertrand, N. Meyer, and M. Maumy-Bertrand. Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. *Bioinformatics*, 31(3):397–404, 10 2014.

[3] P. Bastien and M. Tenenhaus. PLS generalised linear regression, Application to the analysis of life time data In PLS and Related Methods. In *Proceedings of the PLS 01 International Symposium*, page 131–140, Montreuil, 2001. CISIA-CERESTA Editeur.

[4] L. Breiman and P. Spector. Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review*, 60(3):291–319, 1992.

[5] S. L. Brilleman, R. Wolfe, M. Moreno-Betancur, and M. J. Crowther. Simulating survival data using the simsurv r package. *Journal of Statistical Software*, 97(3):1–27, 2021.

[6] S. Cassidy, H. Fuller, J. Chau, M. Catt, A. Bauman, and M.I. Trenell. Accelerometer-derived physical activity in those with cardio-metabolic disease compared to healthy adults: a uk biobank study of 52,556 participants. *Acta Diabetol.*, 55(9):975–979, 2018.

[7] A. Chan, D. Chan, H. Lee, C.C. Ng, and Yeo A.H.L. Reporting adherence, validity and physical activity measures of wearable activity trackers in medical research: A systematic review. *Int J Med Inform.*, 160:104696, 2022.

[8] M. Chen, B. Landré, P. Marques-Vidal, V. T. van Hees, A. C. E. van Gennip, M. Bloomberg, M. S. Yerramalla, M. A. Benadjaoud, and S. Sabia. Identification of physical activity and sedentary behaviour dimensions that predict mortality risk in older adults: Development of a machine learning model in the Whitehall II accelerometer sub-study and external validation in the CoLaus study. *EClinicalMedicine*, 55:101773, 2022.

[9] H. Chun and S. Keles. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. B*, 72:3–25, 2010.

[10] D. Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC, 3 edition, 2014.

[11] I. M. Danilevicz, V. T. van Hees, F. C. T. van der Heide, L. Jacob, B. Landré, M. A. Benadjaoud, and S. Sabia. Measures of fragmentation of rest activity patterns: mathematical properties and interpretability based on accelerometer real life data. *BMC medical research methodology*, 24(1), 2024.

[12] A. Doherty, D. Jackson, N. Hammerla, T. Plotz, P. Olivier, M. H. Granat, T. White, V. T. van Hees, M. I. Trenell, C. G. Owen, S. J. Preece, R. Gillions, S. Sheard, S. Peakman, T.and Brage, and N. J. Wareham. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. *PloS one*, 12(2):e0169649, 2017.

[13] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, page 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[14] K.-A. Lê Cao, S. Boitard, and P. Besse. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 12(253), 2011.

[15] K.-A. Lê Cao and Z.M. Welham. *Multivariate Data Integration Using R: Methods and Applications with the mixOmics Package* . Chapman and Hall/CRC, 1 edition, 2021.

[16] L. Li. Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. *Bioinformatics*, 22(4):466–471, 2006.

[17] T.J. Littlejohns, J. Holliday, L.M. Gibson, S. Garratt, N. Oesingmann, F. Alfaro-Almagro, and et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nat Commun.*, 11(1):2624, 2020.

[18] M. Marmot and E. Brunner. Cohort profile: the Whitehall II study. *Int J Epidemiol.*, 34(2):251–256, 2005.

[19] J.H. Migueles, A.V. Rowlands, F. Huber, S. Sabia, and V.T. van Hees. GGIR: a research community–driven open source R package for generating physical activity and sleep outcomes from multi-day raw accelerometer data. *Journal for the Measurement of Physical Behaviour*, 2(3):188–196, 2019.

[20] S. Nguyen, A. Z. LaCroix, K. M. Hayden, C. Di, P. Palta, M. L. Stefanick, J. E. Manson, S. R. Rapp, M. J. LaMonte, and J. Bellettiere. Accelerometer-measured physical activity and sitting with incident mild cognitive impairment or probable dementia among older women. *Alzheimer's and dementia*, 19(7):3041–3054, 2023.

Declarations

Acknowledges: We thank all of the participating civil service departments and their welfare, personnel, and establishment officers; the British Occupational Health and Safety Agency; the British Council of Civil Service Unions; all participating civil servants in the Whitehall II study; and all members of the Whitehall II study team. The Whitehall II Study team comprises research scientists, statisticians, study coordinators, nurses, data managers, administrative assistants and data entry staff, who make the study possible.

Funding: The Whitehall II study is and was supported by grants from the National Institute on Aging, NIH (R01AG056477, R01AG062553); UK Medical Research Council (R024227, S011676), and the Wellcome Trust (221854/Z/20/Z). SS is funded by the European Union (ERC grant number 101043884). The funding agencies had no role in the study design, data collection, analyses, and interpretation of the data or writing of the manuscript. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the funding agencies. Neither the European Union nor the granting authority can be held responsible for them.

Ethics approval and consent to participate: Written informed consent for participation was obtained at each contact. Research ethics approval was obtained from the University College London ethics committee (latest reference number 85/0938).

Competing interests: The authors declare that they have no competing interests.

Biostatistique 2

Sélection de variables avec contrôle du FDR pour données de grande dimension avec risques concurrents.

Hugo Cannafarina [1] & Guillemette Marot [2] & Evgeniya Babykina [3]

[1] *Modal, Inria Lille et ULR 2694 – METRICS – Évaluation des Technologies de Santé et des Pratiques Médicales, CHU Lille, Université de Lille , France, hugo.cannafarina@inria.fr*
[2] *ULR 2694 – METRICS – Évaluation des Technologies de Santé et des Pratiques Médicales, CHU Lille, Université de Lille, Lille, France, guillemette.marot@univ-lille.fr*
[3] *ULR 2694 – METRICS – Évaluation des Technologies de Santé et des Pratiques Médicales, CHU Lille, Université de Lille, Lille, France, evgeniya.babykina@univ-lille.fr*

Résumé. La méthodologie *Integrated Path Stability Selection* (IPSS), qui permet le contrôle du FDR, est proposée pour la sélection de variables dans une modélisation de type Fine-Gray. Dans le contexte de la grande dimension des prédicteurs, la sélection des variables avec contribution significative à l'estimation de la fonction de survie en présence de risques concurrents reste un défi. Les résultats d'une étude de simulation montrent que la méthodologie développée permet de réduire drastiquement le nombre de faux positifs par rapport aux techniques existantes, tout en conservant une performance robuste dans l'identification des vrais positifs et dans les capacités prédictives. Enfin, la méthode est validée par une étude de cas réelle.

Mots-clés. Risques concurrents, Modèle de Fine-Gray, Sélection de variables, Grande dimension, LASSO, Stability Selection

Abstract : The *Integrated Path Stability Selection* (IPSS) methodology, which enables control of the False Discovery Rate (FDR), is proposed for variable selection in Fine-Gray modeling. In the context of high-dimensional predictors, selecting variables with significant contributions to survival function estimation in the presence of competing risks remains challenging. Simulation study results demonstrate that the developed methodology significantly reduces the number of false positives compared to existing techniques, while maintaining robust performance in identifying true positives and in predictive capabilities. Finally, the method is validated through a real-world case study.

Keywords. Competing risks, Fine-Gray model, Variable Selection, High Dimension, LASSO, Stability Selection

1 Introduction

Dans l'analyse de données de survie, des événements concurrents surviennent souvent avant l'événement d'intérêt, ce qui empêche son observation. Ces événements concurrents doivent être pris en compte lors de l'analyse, car ils affectent la probabilité d'occurrence de l'événement d'intérêt (censure non-indépendente). Les risques concurrents sont particulièrement fréquents dans les études biomédicales. Les sujets peuvent par exemple décéder de plusieurs causes distinctes. Le modèle décrit dans Fine and Gray 1999 est largement utilisé pour évaluer l'impact de covariables sur l'incidence d'un événement en présence de risques concurrents.

Dans les contextes de grande dimension avec plusieurs milliers de variables (données génomiques, transcriptomiques, protéomiques, métabolomiques, *etc.*) et plusiers événements concurrents, la norme des coefficients du modèle de Fine-Gray peut être pénalisée afin de sélectionner les variables associées à un événement d'intérêt. La pénalité LASSO introduite dans Tibshirani 1996 peut par exemple être appliquée en pénalisant la norme L1 des coefficients. Les propriétés théoriques de la version pénalisée du modèle de Fine-Gray ont été étudiées dans Fu, Parikh, and Zhou 2017.

Cependant, les méthodes de régression pénalisés conduisent généralement à l'inclusion de nombreuses variables non pertinentes. En effet, en cas de fortes corrélations, les approches de type LASSO sont connues pour sélectionner une variable au hasard dans un groupe de variables corrélées (Haohan Wang et al. 2019), rendant le processus de sélection très instable. Ces faux positifs entravent les progrès de la recherche biomédicale en compliquant l'identification de biomarqueurs influants.

Plusieurs méthodes alternatives ont été établies pour la régression avec risques concurrents. Binder et al. 2009 a proposé de sélectionner des variables pour le modèle de risque de sous-distributions proportionnel en utilisant un algorithme de boosting, où seuls les coefficients des variables les plus prédictives sont mis à jour à chaque itération. Kawaguchi et al. 2021 a appliqué la régression *Broken Adaptive Ridge*, qui peut être considérée comme un substitut au modèle pénalisé L0, au modèle de risque de sous-distributions proportionnel. Hong Wang et al. 2022 a proposé d'utiliser une méthode de *screening* pour réduire le nombre de faux positifs obtenus lors de l'utilisation des pénalités Lasso et Lasso adaptatif. Sun and X. Wang 2023 a proposé d'augmenter la stabilité du processus de sélection en utilisant un algorithme *split and merge*, qui consiste à appliquer un modèle de Fine-Gray pénalisé par Elastic-Net à des groupes de variable décorrélés.

Nous proposons d'utiliser *Integrated Path Stability Selection* (IPSS, Melikechi and Miller 2024) avec le modèle de Fine-Gray pénalisé. La *Stability Selection*, d'abord présentée dans Meinshausen and Bühlmann 2010, est une technique qui vise à améliorer les performances d'un modèle de sélection de variables en estimant leur probabilité de sélection sur des sous-échantillons aléatoires des données. Elle fournit également une borne sur le nombre moyen de faux positifs. Les performances de la *Stability Selection* ont récemment été améliorées par IPSS, qui offre en plus un contrôle du taux de faux positifs (FDR). Les performances de plusieurs méthodes de sélection de variables dans le contexte des risques concurrents ont déjà été comparées dans des travaux antérieurs (Gilhodes et al. 2017; Hou et al. 2018; Kawaguchi et al. 2021; Li 2023; Mullaert et al. 2024; Tapak et al. 2021). Cependant, à notre

connaissance, les propriétés de la *stability selection* pour le modèle de Fine-Gray n'ont pas encore été étudiées.

À travers des simulations et une étude de cas réelle, nous démontrons que l'utilisation d'IPSS peut réduire le nombre de faux positifs par rapport aux techniques existantes, tout en maintenant de bonnes performances dans l'identification des vrais positifs et la prédiction, soulignant ainsi le potentiel de la méhtode pour améliorer le traitement des données de grande dimension avec risques concurrents dans la recherche biomédicale.

2 Integrated Path Stability Selection pour le modèle de Fine-Gray pénalisé

2.1 Le modèle de Fine-Gray pénalisé

Sans perte de généralité, nous supposerons qu'il existe deux événements possibles, l'événement 1 étant celui d'intérêt et l'événement 2 étant l'événement concurrent. Nous définissons le temps d'évènement T et le temps de censure C. Nous observons $\{Y_i, \delta_i, \delta_i\varepsilon_i, x_i\}$, $i = 1 \dots n$, où $Y_i = \min(T_i, C_i)$, $\varepsilon_i \in \{1, 2\}$ est la variable indicatrice de l'événement, $\delta_i = I(T_i \leq C_i)$ est l'indicateur de non-censure et x_i est un vecteur de covariables de dimension $p \times 1$. On note $\mathbf{x}$ la matrice de covariables de taille $p \times n$ et $F_1(t; \mathbf{x}) = P(T \leq t, \varepsilon = 1|\mathbf{x})$ la Fonction d'Incidence Cumulée (FIC) de l'évènement d'intérêt conditionnellement aux covariables. Pour modéliser les effets des covariables sur la FIC, Fine and Gray 1999 ont proposé le modèle de risque de sous-distribution proportionnel :

$$\begin{aligned}\gamma_1(t; \mathbf{x}) &= dF_1(t; \mathbf{x}) / \{1 - F_1(t; \mathbf{x})\} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P\{t \leq T \leq t + \Delta t, \varepsilon = 1 \mid T \geq t \cup (T \leq t \cap \varepsilon \neq 1), \mathbf{x}\}\end{aligned}$$

Le risque de sous-distribution de l'évènement d'intérêt est supposé suivre un modèle de risque proportionnel log-linéaire, avec $\gamma_1(t; \mathbf{x}) = \gamma_{10} \exp(\beta^\top \mathbf{x})$, où β est un vecteur de covariables de dimension $p \times 1$ et γ_{10} est une fonction de risque de base non spécifiée. La log-vraisemblance partielle du modèle de Fine-Gray est obtenue avec :

$$l(\beta) = \sum_{i=1}^{n} \int_0^\infty \left[\beta^T x_i - \log\left\{\Sigma_j w_j(u) Z_j(u) \exp\left(\beta^T x_j\right)\right\}\right] \times w_i(u) dN_i(u),$$

où $N_i(t) = I\left(T_i \leq t, \varepsilon_i = 1\right)$ est l'indicatrice d'observation de l'événement d'intérêt, $Z_j(t) = 1 - N_j(t-)$ est l'indicatrice d'être à risque pour l'événement d'intéret avant t et $w_i(t) = I\left(C_i \geq T_i \wedge t\right) \hat{G}(t) / \hat{G}\left(T_i \wedge t\right)$ est le poids appliqué aux individus avec $G(t) = P(C \geq t)$ et $\hat{G}(t)$ l'estimation de Kaplan-Meier de C.

Le modèle de Fine-Gray pénalisé est alors obtenu en minimisant $Q(\beta) = -l(\beta) + p_\lambda(\beta)$. Trois pénalités ont été considérées pour l'application de IPSS :

- **LASSO** : $p_\lambda(\beta) = \lambda\|\beta\|_1$
- **LASSO Adaptatif** : $p_\lambda(\beta) = \lambda\|\beta\|_1\theta$, avec θ l'inverse du vecteur des coefficients issu d'une régression Fine-Gray pénalisée par la norme L2.
- **Elastic-Net** : $p_\lambda(\beta) = \frac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1$, avec α fixé à 0.95.

2.2 Integrated Path Stability Selection

La *Stability Selection* permet d'améliorer les capacités des méthodes de sélection de variables en utilisant des sous-échantillons des données. Soit $I \subseteq \{1,\dots,n\}$ un sous-ensemble d'individus. On note $x_I = (x_i : i \in I)$. Nous définissons $\hat{S}_\lambda(x_I) \subseteq \{1,\dots,p\}$ les variables sélectionnées en utilisant le modèle de Fine-Gray pénalisé sur le sous-ensemble I avec le paramètre de pénalité λ définit en section 2.1. Une métrique cruciale de la *Stability Selection* est la probabilité qu'une variable j soit sélectionnée lorsque l'on utilise la moitié du jeu de données. Nous notons cette probabilité de sélection $\pi_j(\lambda)$. Avec IPSS, cette quantité est estimée avec l'algorithme suivant :

Algorithm 1 Probabilités de sélection de variables avec IPSS

1: **Entrée :** Jeu de données $\mathbf{x}$, grille de paramètres Λ, nombre d'itérations B.
2: **Pour** $b = 1,\dots,B$ **faire**
3: Sélectionner aléatoirement des sous-ensembles $I_{2b-1}, I_{2b} \subseteq \{1,\dots,n\}$ avec $I_{2b-1} \cap I_{2b} = \emptyset$ et $|I_{2b-1}| = |I_{2b}| = \lfloor n/2 \rfloor$.
4: **Pour** $\lambda \in \Lambda$ **faire**
5: Calculer $\hat{S}_\lambda\left(x_{I_{2b-1}}\right)$ et $\hat{S}_\lambda\left(x_{I_{2b}}\right)$.
6: **fin Pour**
7: **fin Pour**
8: **Sortie :** Probabilités de sélection $\hat{\pi}_j(\lambda) = \frac{1}{2B}\sum_{b=1}^{2B} I\left(j \in \hat{S}_\lambda\left(x_{I_b}\right)\right)$.

En d'autres termes, $2B$ sous-échantillons du jeu de données, choisis de manière complémentaire deux à deux, sont extraits. Pour chaque valeur de pénalité d'une grille de paramètres donnée, le nombre moyen de fois où chaque variable est sélectionnée eest calculé à travers tous les sous-échantillons, en utilisant le modèle de Fine-Gray avec cette pénalité.

IPSS permet de spécifier un nombre moyen maximum de faux-positifs, noté FP^*. Melikechi and Miller 2024 établissent une formule pour une borne τ_{FP^*} qui dépend de FP^*. Les variables sélectionnées par IPSS, notées $\hat{S}_{IPSS} \subseteq \{1,\dots,p\}$, satisfont :

$$\hat{S}_{\text{IPSS}} = \left\{ j : \int_\Lambda h_m\left(\hat{\pi}_j(\lambda)\right)\mu(d\lambda) \geq \tau_{FP^*} \right\}$$

avec $h_m(x) = (2x-1)^m I(x \geq 0.5), m = 1,\dots,3$ et $\mu(d\lambda) = (\int_\Lambda \gamma^{-1} d\gamma)^{-1}\lambda^{-1}d\lambda$. Sous certaines conditions de régularité, le nombre attendu de variables sélectionnées à tort par IPSS est inférieur à FP^*. Avec des approximations supplémentaires, cette méthode peut également facilement être utilisée pour contrôler le taux de faux positifs.

3 Etude de simulation

Dans le cadre des simulations, la FIC du modèle de Fine-Gray pour l'événement d'intérêt a été fixée à :

$$F_1(t \mid \mathbf{x}) = 1 - \{1 - p\,[1 - \exp(-t)]\}^{\exp(\mathbf{x}^T\beta)}\,,$$

avec $p = 0.75$. Le taux de censure a été fixé à 30% (par le temps de fin de suivi). De plus, 10 variables indexées par $T \subseteq \{1, \ldots, p\}$ ont été choisies comme vrais prédicteurs, avec $\beta_i \sim U(-1, 1), i \in T$. Les autres éléments de β ont été fixées à 0. 100 jeux de données de 5000 variables ont été générés selon le modèle de Toeplitz : $x_i \sim \mathcal{N}(0, \Sigma)$ avec $\Sigma_{jk} = \rho^{|j-k|}$ et $\rho = 0.6$, $i \in 1 \ldots n$. Le nombre d'individus n est fixé à 300.

Les performances de IPSS ont été comparées à celles d'autres méthodes existantes : le modèle Fine-Gray pénalisé avec les pénalités Lasso, SCAD et MCP sur les critères du taux de faux positifs, nombre de vrais-positifs et C-index. Pour IPSS, le taux de faux-positif a été controllé à 0.1 et le paramètre de pénalité a été choisi par BIC pour les autres méthodes. Les résulats sont présentés dans les Figures 1 et 2.

4 Étude de cas réelle

IPSS est appliquée au jeu de données d'une étude sur le cancer de la vessie. La majorité des gènes sélectionnés par la méthode sont soutenus dans la littérature comme étant liés au développement du cancer de la vessie.

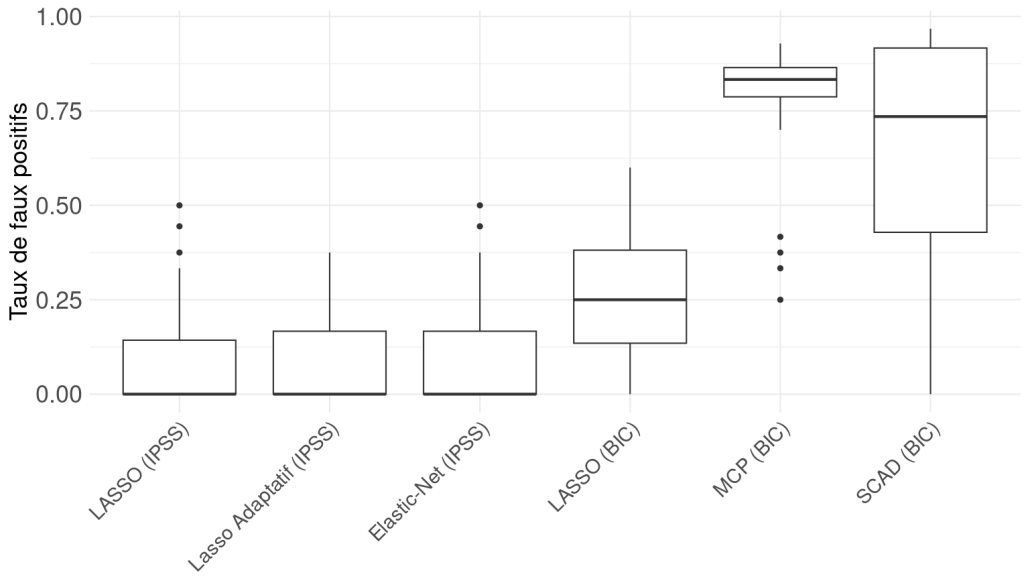


Figure 1: Taux de faux positifs des différents méthodes sur 100 jeux de données de 300 individus, 5000 variables et 10 vrais prédicteurs

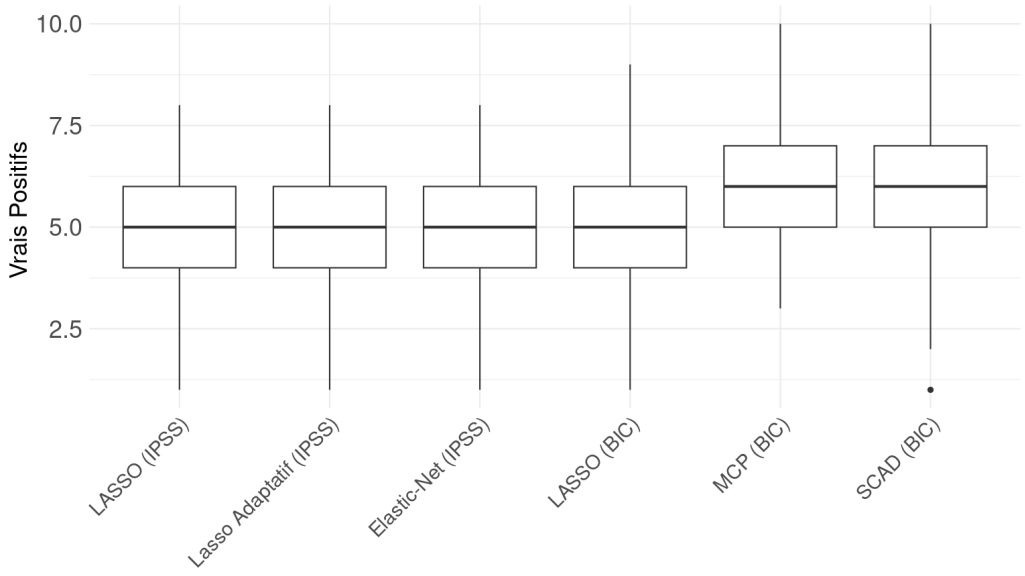


Figure 2: Nombre de vrais positifs des différents méthodes sur 100 jeux de données de 300 individus, 5000 variables et 10 vrais prédicteurs

Bibliographie

Binder, Harald et al. (2009). "Boosting for high-dimensional time-to-event data with competing risks". In: *Bioinformatics* 25.7, pp. 890–896.

Fine, Jason P and Robert J Gray (1999). "A proportional hazards model for the subdistribution of a competing risk". In: *Journal of the American statistical association* 94.446, pp. 496–509.

Fu, Zhixuan, Chirag R Parikh, and Bingqing Zhou (2017). "Penalized variable selection in competing risks regression". In: *Lifetime data analysis* 23, pp. 353–376.

Gilhodes, Julia et al. (2017). "Comparison of variable selection methods for high-dimensional survival data with competing events". In: *Computers in biology and medicine* 91, pp. 159–167.

Hou, Jiayi et al. (2018). "High-dimensional variable selection and prediction under competing risks with application to SEER-Medicare linked data". In: *Statistics in medicine* 37.24, pp. 3486–3502.

Kawaguchi, Eric S et al. (2021). "Scalable algorithms for large competing risks data". In: *Journal of Computational and Graphical Statistics* 30.3, pp. 685–693.

Li, Guowei (2023). *Variable Selection for Competing Risks in High-Dimensional Covariate Spaces Without and With Missing Data*. The Ohio State University.

Meinshausen, Nicolai and Peter Bühlmann (2010). "Stability selection". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72.4, pp. 417–473.

Melikechi, Omar and Jeffrey W Miller (2024). "Integrated path stability selection". In: *arXiv preprint arXiv:2403.15877*.

Mullaert, Jimmy et al. (2024). "Variable selection for competing risk regression models: recommendations for analyzing data from epidemiological studies". In: *medRxiv*, pp. 2024–11.

Sun, Han and Xiaofeng Wang (2023). "High-dimensional feature selection in competing risks modeling: A stable approach using a split-and-merge ensemble algorithm". In: *Biometrical Journal* 65.2, p. 2100164.

Tapak, Leili et al. (2021). "Research Article Regularized Weighted Nonparametric Likelihood Approach for High-Dimension Sparse Subdistribution Hazards Model for Competing Risk Data". In.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1, pp. 267–288.

Wang, Haohan et al. (2019). "Precision Lasso: accounting for correlations and linear dependencies in high-dimensional genomic data". In: *Bioinformatics* 35.7, pp. 1181–1187.

Wang, Hong et al. (2022). "Fast Lasso-type safe screening for Fine-Gray competing risks model with ultrahigh dimensional covariates". In: *Statistics in medicine* 41.24, pp. 4941–4960.

IDENTIFICATION SUPERVISÉE ET NON SUPERVISÉE DE SOUS-GROUPES À L'AIDE D'UN MODÈLE DE PROGRESSION DES MALADIES (Leaspy)

Sofia Kaisaridi [1], Gabrielle Casimiro [1], Juliette Ortholand [1] & Sophie Tezenas du Montcel [1]

[1] *ARAMIS, Sorbonne Université, Institut du Cerveau-Paris Brain Institute-ICM, CNRS, Inria, Inserm, AP-HP, Groupe Hospitalier Sorbonne Université, Paris, France*

sofia.kaisaridi@icm-institute.org, gabrielle.casimiro@icm-institute.org, juliette.ortholand@icm-institute.org, sophie.tezenas@aphp.fr

Résumé. Les modèles de progression des maladies sont des outils prometteurs pour analyser des données longitudinales présentant de multiples modalités. De tels modèles peuvent être utilisés pour estimer la progression de la maladie à long terme et reconstruire des trajectoires individuelles qui peuvent rendre compte de la variabilité entre les patients, mais aussi entre les modalités. Si les modèles classiques modélisent cette variabilité inter-patients comme des perturbations aléatoires autour d'une référence fixe, ceci n'est pas forcément réaliste si des facteurs externes expliquent une partie de la variabilité des trajectoires. Les mutations génétiques, le sexe, les antécédents familiaux, le niveau d'éducation, le statut socio-économique e.t.c. peuvent influencer l'évolution de certaines pathologies. De manière supervisé, nous pouvons répondre au défi de l'identification de sous-groupes d'évolution, par un conditionnement de notre modèle sur les covariables observés à l'inclusion. En revanche, dans certaines maladies chroniques où les mécanismes sous-jacents sont peu étudiés, nous ne disposons pas d'informations suffisantes sur les facteurs externes qui influencent la progression de la maladie. Par conséquent, l'identification de sous-groupes devient un défi plus complexe qui peut être abordé par une approche non-supervisée, basée sur un mélange probabiliste des paramètres du modèle. Dans cette étude nous utilisons un modèle de progression des maladies, non-linéaire à effets mixtes qui passe par une reparamétrisation du temps (*Disease Course Mapping Model* implémenté dans la librairie Python open-source Leaspy) et nous comparons les deux approches. Nous utilisons des donnés de la maladie CADASIL comme notre exemple motivant. Avec le modèle à covariables (approche supervisée) nous allons modéliser le niveau de scores cliniques en fonction de la position de la mutation. Avec le modèle de mélange (approche non-supervisée), différents sous-groupes de progression sont définis à partir uniquement de l' évolution des caractéristiques de la maladie. Une fois que les sous-groupes sont formés nous allons examiner l'effet de la position de la mutation sur leur formation.

Mots-clés. Modélisation de la progression de la maladie, Effets mixtes, MCMC-SAEM, Covariables, Mélange, CADASIL, Mutation

Abstract. Disease progression models are promising tools for analyzing longitudinal data with multiple modalities. Such models can be used to estimate long-term disease progression and reconstruct individual trajectories which can account for variability between patients, but also between modalities. While conventional models model this inter-patient variability as random variations around a fixed reference, this is not necessarily realistic if external factors explain some of the variability in trajectories. Genetic mutations, gender, family history,

level of education, socio-economic status and so on can influence the evolution of certain pathologies. In a supervised way, we can answer the challenge of identifying subgroups of evolution by conditioning our model on the baseline covariates. On the other hand, in certain chronic diseases where the underlying mechanisms are little studied, we do not have sufficient information on the external factors that influence disease progression. As a result, identifying subgroups becomes a more complex task, which can be tackled by an unsupervised approach, based on a probabilistic mixture of model parameters. In this study, we use a non-linear mixed-effects disease progression model that uses time reparametrisation (Disease Course Mapping Model implemented in the Leaspy open-source Python library) and we compare the two approaches. We use data from the CADASIL disease as our motivating example. with the covariate model (supervised approach) we will model the level of clinical scores as a function of mutation position. With the mixture model (unsupervised approach), different progression subgroups are defined solely on the basis of the evolution in disease characteristics. Once the subgroups have been formed, we will examine the effect of the position of the mutation on their formation.

Keywords. Disease Progression Modeling, Mixed-effects, MCMC-SAEM, Covariates, Mixture, CADASIL, Mutation

1 Introduction

Les modèles à effets mixtes sont prometteurs pour analyser des données répétées représentant plusieurs caractéristiques et ils sont souvent utilisés dans les études épidémiologiques. De tels modèles peuvent être utilisés pour estimer une progression à long terme d'une maladie et reconstruire des trajectoires individuelles qui tiennent compte de la variabilité entre les patients (effets aléatoires), mais aussi entre les modalités (effets fixes). Comprendre comment la maladie progresse et quelle est la variabilité attendue entre les individus est essentiel.

Si les modèles classiques modélisent cette variabilité inter-patients comme des perturbations aléatoires autour d'une référence fixe ceci n'est pas forcément réaliste car une partie s'explique parfois par des facteurs externes (et est donc difficilement expliquée uniquement par des perturbations aléatoires). Les mutations génétiques ou des facteurs externes tels que le sexe, les antécédents familiaux, le niveau d'éducation ou le statut socio-économique peuvent influencer l'évolution de certaines pathologies. D'un autre point de vue, dans certaines maladies chroniques où les mécanismes sous-jacents sont peu étudiés, nous disposons de peu ou pas d'intuition sur les covariables à l'inclusion qui influencent la progression de la maladie. Par conséquent, l'identification de sous-groupes devient un défi plus complexe, pouvant être abordé par un mélange probabiliste des paramètres du modèle.

Dans ce cadre, les modèles de progression de maladie (*Disease Progression Models*) sont des outils émergents qui reconstruisent les chronologies des maladies chroniques à long terme, fournissant ainsi un aperçu unique des processus pathologiques et de leurs mécanismes sous-jacents (Young et al. (2024)). Ils sont interprétables, facilitant ainsi la compréhension des maladies. Dans cette étude nous avons utilisé un modèle de progression des maladies, non-linéaire à effets mixtes, qui passe par une reparamétrisation du temps, appelé *Disease Course Mapping Model*, comme proposé par Schirrati et al. (2017). Les deux approches décrites ci-

dessus peuvent être considérées comme des extensions du modèle DCM, correspondant respectivement à un modèle tenant compte des covariables et à un modèle de mélange gaussien. Nous étendons le modèle de mélange existant pour inclure tous les aspects de la variabilité spatio-temporelle. L'objectif de ce travail est d'éclairer la manière dont ces deux techniques peuvent être modélisées dans le cadre du modèle DCM et de comparer leurs résultats. Les deux répondent à la même problématique : l'identification de différents sous-groupes de progression de la maladie ainsi que les facteurs qui les caractérisent. Nous utilisons comme exemple motivant les données de la cohorte de patients atteints de CADASIL recrutés au Centre national de référence français CERVCO.

2 Disease Course Mapping Model

2.1 Modèle générique

Nous considérons des données longitudinales, c'est-à-dire des patients avec des mesures répétées dans le temps. Le modèle de progression utilisé, c'est-à-dire le modèle DCM implémenté dans une librairie Python open-source (Leaspy), a déjà fait ses preuves pour décrire la progression des maladies de façon multivarié, et a déjà été appliqué aux diverses maladies neurodégénératives (Kaisaridi et al. (2025), Ortholand et al. (2023), Maheux et al. (2023), Poulet et Durrleman (2023), Moulaire et al. (2023) Koval et al. (2022)).

Nous supposons un ensemble de données longitudinales $(y_{ijk})_{1\leq i\leq n, 1\leq j\leq N_i, 1\leq k\leq d}$ qui décrivent les valeurs pour chaque patient i à la visite j de la caractéristique k. Le nombre total de visites N_i peut être différent pour chaque patient i. Les données y_{ijk} sont supposées être des points sur une variété riemannienne $\mathcal{M}$. Les paramètres de population (effets fixes) décrivent la trajectoire moyenne comme une géodésique γ_0 sur la $\mathcal{M}$ avec $\gamma_0(t_0) = \mathbf{p}$ et $\dot{\gamma_0}(t_0) = \mathbf{v}$. Ici nous utilisons la forme logistique du modèle :

$$y_{ijk} = (1 + (\frac{1}{p_k} - 1)exp(-\frac{v_k(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0) + w_{ik}}{p_k(1 - p_k)}))^{-1} + \epsilon_{ijk} \tag{1}$$

avec une erreur résiduelle $\epsilon_{ij} \sim \mathcal{N}(\mathbf{O}_d, \sigma_k^2 \mathbf{I}_d)$. Pour simplifier, notre modèle construit la trajectoire moyenne de chaque score sous la forme d'une courbe logistique décrite par les paramètres p_0, v_0 et t_0, où p_0 et v_0 sont la position et la vitesse (dérivée de la courbe) au temps t_0, le point médian de la logistique. La reparamétrisation du temps individuel prend la forme suivante : $\psi_i(t) = \alpha_i(t - t_0 - \tau_i) + t_0$ où τ_i est le décalage temporel et $\alpha_i = e^{\xi_i}$ est le facteur d'accélération. Le profil temporel des patients est décrit par deux paramètres : le décalage temporel τ_i qui correspond à l'âge estimé du début de la maladie, mesuré en années, et le taux de progression ξ_i qui indique si la progression globale est accélérée (valeurs positives) ou ralentie (valeurs négatives). Le profil spatial est défini par les paramètres d'espacement intermarqueurs ω_{ik} qui tiennent compte de l'ordre différent des événements au sein de notre population. Pour chaque patient, nous avons un ω pour chaque score clinique inclus dans le modèle avec des valeurs négatives indiquant qu'un score spécifique commence à se détériorer plus tôt pour un patient spécifique que pour la population moyenne. Néanmoins pour une interprétabilité accrue, le modèle utilise une analyse en composantes indépendantes avec N_s

sources indépendantes, ce qui conduit à des paramètres de décalage spatial $\mathbf{w}_i = As_i$ tels que les colonnes $\mathbf{A}_{:,l} = \sum_{m=1}^{d-1} \beta_{ml}\mathcal{B}_m$ sont une combinaison linéaire d'une base orthonormale $(\mathcal{B}_m)_{1\leq m\leq d-1}$ du sous-espace orthogonal à $Span(\mathbf{v})$. Le modèle statistique hiérarchique suppose que les paramètres de population $\mathbf{z}_{pop} = (\tilde{g_k}, \tilde{v_k}, \beta_{ml})$ et les paramètres individuels $z_i = (\xi_i, \tau_i, s_{il})$ sont latents et suivent des distributions gaussiennes :

$$\begin{cases} \xi_i \sim \mathcal{N}(0, \sigma_\xi^2) \\ \tau_i \sim \mathcal{N}(\overline{\tau}, \sigma_\tau^2) \\ s_{il} \sim \mathcal{N}(0, 1) \end{cases} \quad (2) \qquad \begin{cases} \tilde{g_k} \sim \mathcal{N}(\overline{\tilde{g_k}}, \sigma_{\tilde{g}}^2) \\ \tilde{v_k} \sim \mathcal{N}(\overline{\tilde{v_k}}, \sigma_{\tilde{v}}^2) \\ \beta_{ml} \sim \mathcal{N}(\overline{\beta_{ml}}, \sigma_\beta^2) \end{cases} \quad (3)$$

Enfin les paramètres du modèle statistique sont $\theta = (\overline{\tilde{g}}_k, \overline{\tilde{v}}_k, \overline{\beta}_{ml}, \overline{\tau}, \sigma_\tau, \sigma_\xi, \sigma_k)$ tandis que $\sigma_{\tilde{g}}$, $\sigma_{\tilde{v}}$, σ_β sont fixes.

2.2 Deux approches pour l'identification de sous-groupes

Le modèle à covariables définit les sous-groupes de manière déterministe et alors, il est susceptible de nous fournir des estimations plus stables et plus précises sur l'évolution de la maladie dans les différents groupes. En revanche, il exige une connaissance préalable des facteurs qui affectent la progression de la maladie, ce qui n'est pas toujours le cas. De l'autre côté le modèle de mélange, ne nécessitant aucune connaissance a priori, a un champ d'application plus large et conclut à l'identification de sous-types présentant des profils de progression similaires, après avoir examiné une série de répartitions possibles. Pourtant sa nature probabiliste s'accompagne d'un coût computationnel plus élevé, et pourrait nous amener à une convergence plus lente. Nous allons décrire les deux méthodes ci-dessous.

2.2.1 Approche supervisée : Modèle à covariables

Au lieu d'estimer des effets fixes (paramétrant l'évolution moyenne de la maladie) ainsi que des effets aléatoires, nous introduisons une fonction lien f_ϕ capable de prédire, à partir d'un ensemble de covariables c_i, une trajectoire attendue de la maladie conditionnée par ces covariables (Fournier et Durrleman (2023)). La principale différence par rapport à l'approche standard réside dans le fait que les effets fixes, précédemment introduits, sont désormais estimés comme une fonction déterministe de covariables $f_\phi(c_i)$. Dans ce cas, les covariables sont utilisées de manière supervisée lors de la calibration du modèle et servent à naviguer à travers un continuum de modèles de maladies, au lieu d'avoir des clusters définis.Nous paramétrons la fonction lien f_ϕ comme une correspondance linéaire entre les covariables et les paramètres de population :

$$f_{\boldsymbol{\phi}}(c_i) = \boldsymbol{\phi}_{slope} \cdot c_i + \boldsymbol{\phi}_{intercept}$$

C'est-à-dire pour chacun des paramètres $(\tilde{g_k}, \tilde{v_k}, , \overline{\tau})$ on calcule une paire de paramètres $\boldsymbol{\phi}$ correspondant à la pente et à l'intercept de la formulation linéaire. Ces paramètres $\boldsymbol{\phi}$ nous donnerons les valeurs de $(\tilde{g_k}, \tilde{v_k}, , \overline{\tau})$ pour chaque valeur possible des covariables c_i. Ainsi, les paramètres du modèle statistique sont $\theta = (\bar{\boldsymbol{\phi}}, \bar{\beta}_{ml}, \sigma_\xi, \sigma_\tau, \sigma)$ tandis que $\sigma_{\boldsymbol{\phi}}$, σ_β sont fixes.

2.2.2 Approche non-supervisée : Modèle de mélange

Pour créer un modèle de mélange, nous ajoutons une nouvelle couche au-dessus de la structure hiérarchique déjà construite et nous nous inspirons de la méthode décrite dans Poulet et Durrleman (2021). Notre contribution réside dans le fait que nous intégrons les paramètres individuels spatials (les sources s_{il}) dans le mélange probabiliste. Nous supposons que chaque cluster c est défini par ses paramètres $(\overline{\xi}^c, \sigma_\xi^c, \overline{\tau}^c, \sigma_\tau^c, \overline{s}^c)$ et décrit ainsi un sous-type avec une moyenne et un écart-type différents pour le début de la maladie et le rythme de progression, mais aussi des moyennes différentes pour les paramètres contrôlant l'ordre d'apparition des caractéristiques mesurées. Ainsi, les clusters formés se différencient à la fois en termes de progression spatiale et temporelle. Dans ce cadre les paramètres individuels suivent des distributions de mélange de gaussienne et donc l'équation 2 est remplacé par :

$$\begin{cases} \xi_i \sim \sum_{c=1}^{n_c} \pi^c \mathcal{N}(\overline{\xi}^c, \sigma_\xi^c), & with \quad \sum_{c=1}^{n_c} \pi^c \overline{\xi}^c = 0 \\ \tau_i \sim \sum_{c=1}^{n_c} \pi^c \mathcal{N}(\overline{\tau}^c, \sigma_\tau^c) & \\ s_{il} \sim \sum_{c=1}^{n_c} \pi^c \mathcal{N}(\overline{s}^c, 1), & with \quad \sum_{c=1}^{n_c} \pi^c \overline{s}^c = 0 \end{cases} \tag{4}$$

avec π^c les coefficients de mélange qui correspondent aux probabilités d'occurrence de chaque cluster c. Par conséquent, nous considérons que chaque patient appartient au cluster c, avec une probabilité individuelle π_i^c dépendant de la vraisemblance des effets aléatoires. L'estimation conjointe des clusters et des paramètres du modèle est effectuée à l'aide d'un algorithme de mélange MCMC avec approximation stochastique (M-MCMC SAEM).

3 Cadre d'application

L'artériopathie cérébrale autosomique dominante à l'origine d'infarctus sous-corticaux et d'une leucoencéphalopathie (CADASIL) est la maladie héréditaire la plus courante touchant les petits vaisseaux cérébraux (SVD) et elle est causée par des mutations du gène NOTCH3 (Rutten et al. (2020)). Les patients atteints de CADASIL présentent un large éventail de symptômes tels que des crises de migraines avec aura, des accidents vasculaires cérébraux, des troubles du comportement, une incapacité motrice et des troubles cognitifs de la dysfonction exécutive jusqu'à une démence sévère (Chabriat et al. (2009)). L'étude de sous-groupes de patients CADASIL (Kaisaridi et al.(2025)) a montré que les patients présentant un début précoce et un rythme de progression rapide développent plus tôt des symptômes d'incapacité, de dépendance et de déficits focaux (éventuellement liés aux accidents vasculaires cérébraux). De l'autre côté les patients présentant un début plus tardif et un rythme de progression plus doux, développent d'abord des symptômes cognitifs. Globalement, le sexe, le niveau d'éducation, la position de la mutation dans les domaines de l'EGFr et certains facteurs de risque cardiovasculaire peuvent affecter la progression clinique de la maladie.

Avec le modéle à covariables nous allons évaluer l'intensité de scores cliniques en fonction de la position de la mutation de maniére supervisée. Nous supposons à priori que la position de la mutation est un facteur de risque qui affecte la progression de la maladie. De l'autre côté avec le modéle de mélange que nous avons implémenté dans Leaspy nous allons d'abord identifier des sous-groupes d'évolution et ensuite explorer l'effet de la position de la

mutation. Nous adoptons une approche non-supervisé où les différents profils de progression sont basés uniquement sur l' évolution des caractéristiques de la maladie. Une fois que les sous-groupes sont formés nous allons examiner l'effet de la position de la mutation sur leur formation. Nous comparerons ensuite les sous-groupes issus de l'approche supervisée et ceux obtenus par l'approche non-supervisée afin d'évaluer leur pertinence. Cette comparaison nous permettra de mieux comprendre dans quelle mesure la position de la mutation influence la progression de la maladie et si les sous-groupes identifiés de manière non-supervisée capturent des dynamiques similaires ou distincte.

Bibliographie

Chabriat, H., Joutel, A., Dichgans, M., Tournier-Lasserve, E. et Bousser, M.G. (2009), Cadasil. *The Lancet. Neurology*, 8(7), pp. 643-653.

Fournier N., Durrleman S. (2023). A Multimodal Disease Progression Model for Genetic Associations with Disease Dynamics. *Medical Image Computing and Computer Assisted Intervention, MICCAI, Oct 2023, Vancouver, Canada*, pp.601-610

Kaisaridi S., Herve D., Jabouley A., Reyes S., Machado C., Guey S., Taleb A., Fernandes F., Chabriat H. et Tezenas du Montcel S. (2025). Determining Clinical Disease Progression in Symptomatic Patients With CADASIL. *Neurology*, 104(1), e210193

Koval I., Dighiero-Brecht T., Tobin A. J., Tabrizi S. J., Scahill R. I., Tezenas du Montcel S., Durrleman S., et Durr A. (2022), Forecasting individual progression trajectories in Huntington disease enables more powered clinical trials. *Scientific reports - Nature*, 12(1), 18928.

Maheux E., Koval I., Ortholand J., Birkenbihl C., Archetti D., Bouteloup V., Epelbaum S., Dufouil C., Hofmann-Apitius M., Durrleman S. (2023). Forecasting individual progression trajectories in Alzheimer's disease. *Nature Commmunications*, 14(1), pp. 761

Moulaire P., Poulet P.E., Petit E., Klockgether T., Durr A., Ashisawa T., Tezenas du Montcel S., for the READISCA Consortium (2023). *Movement Disorders*, 38(1), pp. 35-44

Ortholand J., Pradat P.F., Tezenas du Montcel S., et Durrleman S. (2023). Interaction of sex and onset site on the disease trajectory of amyotrophic lateral sclerosis. *Journal of Neurology*, 270(12), pp. 5903-5912.

Poulet P. E., Durrleman S. (2023). Multivariate disease progression modeling with longitudinal ordinal data. *Statistics in Medicine*, 42(18), pp. 3164-3183

Poulet P. E., Durrleman S. (2021). Mixture modeling for identifying subtypes in disease course mapping. *Information Processing in Medical Imaging, 27th International Conference*

Rutten, J. W., Hack, R. J., Duering, M., Gravesteijn, G., Dauwerse, J. G., Overzier, M., van den Akker, E. B., Slagboom, E., Holstege, H., Nho, K., Saykin, A., Dichgans, M., Malik, R., et Lesnik Oberstein, S. A. J. (2020), Broad phenotype of cysteine-altering NOTCH3 variants in UK Biobank : CADASIL to nonpenetrance. *Neurology*, 95(13), pp. e1835-e1843.

Schirratti J.B., Allassonniére S., Colliot O., et Durrleman S. (2017), A Bayesian Mixed- Effects Model to Learn Trajectories of Changes from Repeated Manifold-Valued Observations. *Journal of Machine Learning Research*, 18, pp. 1-33.

Young, A. L., Oxtoby, N. P., Garbarino, S., Fox, N. C., Barkhof, F., Schott, J. M., et Alexander, D. C. (2024), Data-driven modelling of neurodegenerative disease progression : thinking outside the black box. *Nature reviews.* Neuroscience, 25(2), pp. 111-130.

Testing for misclassified deaths in cancer registries by comparing relative and cause-specific survival

Oskar Laverny [1]

[1] *Aix Marseille Univ, INSERM, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Marseille, France.*
oskar.laverny@univ-amu.fr

Résumé. Le reporting de la cause de la mort dans les registres de cancer est un point de discussion historique. Son manque de qualité voir son absence totale empêche l'utilisation de modèles de risques compétitifs standard. Pour contourner le problème, les méthodes de survie relative utilisent, outre quelques hypothèses, la connaissance des tables de mortalité populationnelles "toutes causes", pour estimer néanmoins la mortalité vis-à-vis du cancer. La littérature contient plusieurs vérifications empiriques de la qualité de la cause de mort rapportée sur de vraies cohortes tirées des registres, mais aucune n'est sérieusement formalisée. Nous proposons pour la première fois un test statistique formel pour vérifier la qualité de la cause de mort rapportée, en exploitant les techniques de survie relative d'un côté et de risques compétitifs de l'autre. Notre test peut être utilisé par des praticiens pour vérifier la cohérence de la cause de la mort rapportée dans leurs jeux de données, sous les hypothèses standard de la survie nette. Nous étudions le comportement et la calibration de ce test sur données simulées, avant de fournir une illustration sur des données réelles tirées du registre Américain SEER.

Mots-clés. Analyse de survie, Survie relative, estimateurs non paramétriques, tests, ...

Abstract. Historically, the cause of death reported in cancer registries was often unreliable, and this crucial information could not be exploited by competing-risks methods. To circumvent this issue, the relative survival methodology leveraged the knowledge of general population mortality from national life tables, together with a few assumptions, to still estimate the cancer-specific mortality. The literature contains several empirical studies of the quality of the cause-of-death reported in real cancer registries datasets, but none are seriously formalized. We propose for the first time a formal statistical test for the quality of the cause of death reported in the databases, exploiting techniques from both the relative survival and competing risks methodologies. Our test can be used by practitioners to assess the alignment of their reported causes of death, under the standard relative survival hypotheses. We study the behavior and calibration of the tests on simulated datasets and finally provide an illustration using the SEER registry.

Keywords. Survival analysis, Relative survival, Non-parametric estimators, Tests, ...

1 Formal context

The relative survival methodology deals with a competing risks setup where there are only two potential causes of death: the studied cancer or anything else, and where the reported cause of death is unreliable. This particularity is taken into account to evaluate the excess mortality – due to cancer – with respect to general population life tables.

More formally, let E, P and $O = E \wedge P$ be random times to death w.r.t. (resp) the Excess, Population and Overall mortalities. Let $\mathbf{X}$ be a vector of covariates, C the time to censorship, and denote $T = O \wedge C$ the time to any event, $\Delta = \mathbf{1}\{T \leq C\}$ the censorship indicator and $\Gamma = \mathbf{1}\{E \leq P\}$ the cause of death indicator.

The cause of death indicator Γ is the problematic one, often misreported in cancer registries. Historically, the relative survival methodology thus did not use this information at all to estimate the excess mortality. The fiability of the cause of death was discussed extensively in the literature (see e.g. Iczi & al (2021) for a recent discussion), but never formally. The objective of this work is to produce a formal testing methodology to verify if this indicator is, indeed, misreported or not.

We thus observe $(\mathbf{X}, T, \Delta, \Gamma)$, with a doubt on the distribution of Γ. As in standard relative survival, we suppose the distribution of $P|\mathbf{X}$ to be known from population life tables. To simplify our exposition, assume that everything else does not depend on covariates, and moreover that $E, (P, \mathbf{X})$ and C are mutually independent.

Let us define the following stochastic processes:

$$
\begin{aligned}
N(t) &= \mathbb{1}\{O \leq t, O \leq C\} && \textit{(Uncensored deaths process)}\\
Y(t) &= \mathbb{1}\{O \geq t, C \geq t\} && \textit{(At-risk process)}\\
M(t) &= N(t) - \int_0^t Y(s)\partial\Lambda_O(s) && \textit{(Martingale)}\\
N_E(t) &= \mathbb{1}\{E \leq t, E \leq C\} && \textit{(Excess uncensored deaths process)}\\
Y_E(t) &= \mathbb{1}\{E \geq t, C \geq t\} && \textit{(Excess at risk process)}\\
\Gamma(t) &= \mathbb{1}\{E = t, E \leq P\}\\
N^e(t) &= \mathbb{1}\{E \leq t, E \leq P \wedge C\}\\
N^p(t) &= \mathbb{1}\{P \leq t, P \leq E \wedge C\},
\end{aligned}
$$

and remark that $N^e = \Gamma N$, $N^p = (1 - \Gamma)N$ and finally that $N = N^e + N^p$.

Let $(\mathbf{X}_i, T_i, \Delta_i, \Gamma_i)_{i=1,...,n}$ be an observed n-sample, N_i, Y_i, M_i, N_i^e and N_i^p the associated processes and $(\Omega, \mathcal{A}, \{\mathcal{F}_t, t \in \mathbb{R}_+\}, \mathbb{P})$ the associated filtered probability space, with

$$\mathcal{F}_t = \sigma\left\{\mathbf{X}_i, (T_i, \Delta_i, \Gamma_i) : T_i \leq t,\ \forall i \in 1, .., n\right\}.$$

Let us finally denote for any random variable X its survival function $S_X(t) = \mathbb{P}(X > t)$ and its cumulative hazard function $\Lambda_X(t) = -\ln S_X(t)$.

Our null hypothesis is that $\Gamma_1, ..., \Gamma_n$ are indeed correctly reported, that is :

$$(\mathcal{H}_0): \quad \forall i \in 1, ..., n,\ \Gamma_i = \mathbb{1}\{E_i \leq P_i\}.$$

2 Proposed test

Relative survival assumes the mutual independence of (E, P, C), and, without observing Γ, estimates the distribution of E by the Pohar Perme (2012) estimator:

$$\partial\widehat{\Lambda_E}(t) = \frac{\sum_{i=1}^n w_i(t)\partial N_i(t) - w_i(t)Y_i(t)\partial\Lambda_{P_i}(t)}{\sum_{i=1}^n w_i(t)Y_i(t)},$$

where weights $w_i(t)$ are defined as $w_i(t) = S_{P_i}(t)^{-1}$. This estimator is unbiased, and moreover $\widehat{\Lambda_E} - \Lambda_E$ is a martingale (see Laverny, Grafféo & Giorgi (2025) for its construction). On the other hand, assuming that Γ is reliable, due to the independence, through similar arguments we propose to estimate Λ_E by the following Kaplan-Meier-like estimator:

$$\partial\widetilde{\Lambda_E}(t) = \frac{\sum_{i=1}^n w_i(t)\partial N_i^e(t)}{\sum_{i=1}^n w_i(t)Y_i(t)}.$$

Other estimators could be constructed with weighting schemes alike Wells, Rutherford & Lambert (2024), but we show that $\widetilde{\Lambda_E}(t)$ is unbiaised and has Doob-Meier decomposition $\widetilde{\Lambda_E}(t) = \Lambda_E(t) + \widetilde{\Xi}(t)$, where:

$$\partial\widetilde{\Xi}(t) = \frac{\sum_{i=1}^n \partial w_i(t)M_i^e(t)}{\sum_{i=1}^n w_i(t)Y_i(t)}.$$

A natural choice for our test is therefore to compare these two estimators, and it yields

$$Z(t) = \widehat{\Lambda_E}(t) - \widetilde{\Lambda_E}(t) = \int_0^t \frac{\sum_{i=1}^n w_i(s)\partial M_i^p(s)}{\sum_{i=1}^n w_i(s)Y_i(s)},$$

where $M_i^p(t)$ is the martingale associated with the populational deaths process N_i^p. Under the null, we show that Z is a centered locally square integrable martingale, with optional process

$$\partial\left[Z\right](t) = \frac{\sum_{i=1}^n w_i(t)^2\partial N_i^p(t)}{\left(\sum_{i=1}^n w_i(t)Y_i(t)\right)^2}.$$

By standard arguments, Z is asymptotically Gaussian, and we construct our test on the statistic $\widehat{\chi^2}(t)$ defined as

$$\widehat{\chi^2}(t) = \frac{Z(t)^2}{\left[Z\right](t)} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(1).$$

This result gives us the wanted test procedure: if $\Gamma_1, ...\Gamma_n$ are misreported, $\widehat{\chi^2}(t)$ will deviate from the $\chi^2(1)$ distribution. Moreover, the expression we obtained for Z is particularly readable, as the difference between an empirical and a theoretical value for the marginal populational hazard.

Extension to non-proportional divergences are on the table through over Omnibus tests (such as the one from Wolski & al (2020)).

3 Applications

We propose to assess the performance of our tests on simulated data first, to ensure correct calibration under serveral hypothesis for the bias of Γ_i's, including the null. We finally illustrate the approach on colorectal cancer data from the Surveillance, Epidemiology, and End Results (SEER) database as in e.g. Czado & Van Keilegom (2023). The technical code providing the test will at term be hosted in the Julia package `NetSurvival.jl` from Alhajal & Laverny (2024).

Bibliographie

Pohar Perme, M., Stare, J., and Estève, J. (2012), On estimation in relative survival. In: *Biometrics*, vol. 68, no 1, p. 113-120

Laverny, O., Grafféo, N., and Giorgi, R. (2025), Non-parametric estimation of net survival under dependence between death causes. In: *arXiv preprint*.

Wells, M., Rutherford, M. J., and Lambert, P. C. (2023), Fair comparisons of cause-specific and relative survival by accounting for the systematic removal of patients from risk-sets. In: *Cancer Epidemiology*, 86, 102408.

Wolski, A., Graffeo, N., Giorgi, R., and CENSUR working survival group. (2020), A permutation test based on the restricted mean survival time for comparison of net survival distributions in non-proportional excess hazard settings. In: *Statistical Methods in Medical Research*, 29(6), 1612-1623.

Czado, C. and Van Keilegom, I. (2023), Dependent censoring based on parametric copulas. In: *Biometrika* 110(3), 721-738.

Alhajal, R., and Laverny, O. (2024), `NetSurvival.jl`: A glimpse into relative survival analysis with Julia. In: *arXiv preprint* arXiv:2408.15655.

Izci, H., Tambuyzer, T., Vandeven, J., Xicluna, J., Wildiers, H., Punie, K., ... & Verdoodt, F. (2021). Cause of death for patients with breast cancer: discordance between death certificates and medical files, and impact on survival estimates. In: *Archives of Public Health*, 79(1), 111.

Classification

Vitesses rapides en classification dans le modèle logistique: performance de l'estimateur du maximum de vraisemblance au delà du lemme de Zhang

Hugo Chardon [1] , Matthieu Lerasle [2] & Jaouad Mourtada [3]

[1] *CREST - ENSAE - IP Paris, France hugo.chardon@ensae.fr*
[2] *CREST - ENSAE - IP Paris, France matthieu.lerasle@ensae.fr*
[3] *CREST - ENSAE - IP Paris, France jaouad.mourtada@ensae.fr*

Résumé. Nous nous intéressons ici au problème classique de classification binaire en apprentissage supervisé. La minimisation du risque empirique pour la perte usuelle de classification est un problème non convexe donc difficilement traitable numériquement. Pour contourner ce problme, une idée naturelle est d'utiliser des substituts convexes àu lieu de la perte binaire. L'enjeu est alors de garantir que les prédicteurs obtenus par minimisation du substitut convexe du risque ont aussi de bonnes performances pour le problème initial de classification binaire. Le résultat fondateur de Zhang (2004) montre que ceci est vrai pour une large collection de substituts convexes, en particulier le risque logistique. Le prix à payer dans l'utilisation de ce résultat est une dégradation de la vitesse de convergence vers 0 de l'excès de risque. En effet, le lemme de Zhang appliqué à la perte logistique montre que l'excès de risque en classification est majoré par *la racine carrée de* l'excès de risque logistique. Nous montrons ici que la perte de cet exposant 1/2 n'est pas nécessaire.

Mots-clés. Apprentissage statistique, régression logistique, grande dimension, analyse non asymptotique, processus empiriques, classification, lemme de Zhang, substituts convexes.

Abstract. We are here interested in the classical problem of supervised binary classification. Empirical risk minimisation for the usual binary loss is a computationally intractable problem. A natural idea to circumvent this issue is to replace the binary loss function by a convex surrogate. It is therefore necessary to show that minimisers of the surrogate risk also have a low binary excess risk to guarantee good performances in the original problem. The seminal result of Zhang (2004) shows that this is the case for a wide range of convex surrogates. The general consistency guarantee offered by Zhang's lemma comes at the price of a degraded rate of convergence of the excess risk. Namely, for the logistic risk, one loses a power of 1/2 between the excess surrogate risk and the excess binary risk. We show that this degradation can be avoided.

Keywords. Statistical learning, logistic regression, high dimension, finite-sample analysis, empirical processes, classification, Zhang's lemma, convex surrogates.

1 Introduction

Considérons le problème de classification binaire, dans lequel nous cherchons à prédire une étiquette $y \in \{-1, 1\}$ à partir de covariables $x \in \mathcal{X}$ (nous considérerons principalement le cas où $\mathcal{X}$ est l'espace euclidien $\mathbb{R}^d$, $d \geqslant 1$). Dans le cadre de l'apprentissage statistique, on considère le couple (x, y) comme la réalisation d'une variable aléatoire de loi jointe inconnue P sur $\mathcal{X} \times \{-1, 1\}$. En classification, la fonction de perte naturelle pour mesurer la qualité d'une prédiction est celle qui renvoie 1 si cette dernière est fausse et 0 sinon. La qualité d'un prédicteur $f : \mathcal{X} \to \{-1, 1\}$ est quant à elle mesurée par son risque $L(f)$ qui est l'erreur moyenne par rapport à la loi P, c'est-à-dire

$$L(f) = \mathbb{E}[\mathbf{1}(f(X) \neq Y)] = \mathbb{P}(Yf(X) < 0)\,. \tag{1.1}$$

Comme la loi jointe P est inconnue, la fonction de risque L l'est aussi. On dispose en revanche d'un jeu de données $(X_1, Y_1), \ldots, (X_n, Y_n)$ qui sont des copies indépendantes de (X, Y). La tâche d'apprentissage consiste à utiliser ces données pour construire un prédicteur $\widehat{f}_n$ ayant un risque aussi faible que possible. On s'intéresse plus spécifiquement ici aux règles d'apprentissage par minimisation du risque empirique, qui sélectionnent les prédicteurs faisant la plus petite erreur moyenne sur le jeu de données, c'est-à-dire les solutions du problème

$$\min_{f \in \mathcal{F}} \widehat{L}_n(f)\,, \quad \widehat{L}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(Y_i f(X_i) < 0)\,, \tag{1.2}$$

où $\mathcal{F}$ est un sous-ensemble des fonctions mesurables $\mathcal{X} \to \{-1, 1\}$, appelé classe de comparaison. Cette approche est au cœur de la théorie de Vapnik et Chervonenkis (Vapnik et Chervonenkis (1974); Vapnik (2000)), et basée sur l'heuristique de la convergence uniforme du risque empirique $\widehat{L}_n$ vers son espérance L. Contrairement à L, le risque empirique $\widehat{L}_n$ peut être évalué. Cependant, il s'agit d'une fonction non convexe (en f), rendant le problème (1.2) numériquement inabordable. Afin de tirer parti de l'efficacité des algorithmes d'optimisation convexe, une idée naturelle a été de remplacer la perte non convexe $\ell(f; (X, Y)) = \mathbf{1}(Yf(X) < 0)$ par un substitut convexe $\ell_\phi(f; (X, Y)) = \phi(-Yf(X))$, où ϕ est une fonction convexe satisfaisant certaines propriétés, dites de *calibration*, introduites par Bartlett *et al.* (2006). On s'intéresse dès lors aux prédicteurs ayant un faible ϕ-risque $L_\phi(f) = \mathbb{E}[\phi(-Yf(X))]$. Ceux-ci sont obtenus par minimisation du ϕ-risque empirique puisque la loi jointe du couple (X, Y) est toujours inconnue.

L'enjeu clé du point de vue statistique est d'assurer que les solutions du problème associé au substitut convexe ont aussi de bonnes propriétés de prédiction. Ceci est l'objet du résultat central de Zhang (2004), qui montre que si ϕ est un substitut convexe calibré, une règle d'apprentissage dont le ϕ-risque tend vers 0 a aussi un excès de risque en classification tendant vers 0. Plus précisément, le lemme de Zhang montre que sous certaines conditions sur la loi jointe P (en particulier sur le comportement de la *fonction de régression* $\eta(X) = \mathbb{P}(Y = 1|X)$), on a pour tout $g : \mathcal{X} \to \mathbb{R}$ dans $L^2(P_X)$

$$L(\mathrm{sign}(g)) - L(f^*) \leqslant \Psi\big(L_\phi(g) - L_\phi(f^*)\big)\,, \tag{1.3}$$

où Ψ est une fonction concave. Ici, L désigne le risque binaire (1.1) et $\mathrm{sign}(t) = \mathbf{1}(t > 0) - \mathbf{1}(t \leqslant 0)$ la fonction signe. En particulier, si $\phi(t) = \log(1 + e^t)$ est le risque logistique, alors (1.3) est satisfaite avec $\Psi(t) \asymp \sqrt{t}$.

Notons que dans plusieurs cas d'application, la perte convexe associée à ϕ ne résulte pas seulement d'une astuce numérique dont il faut simplement prouver le bien-fondé statistique avec une borne de type (1.3). Le risque logistique, notamment, apparaît naturellement dans l'estimation par maximum de vraisemblance en régression logistique et est porteur d'une information sur la qualité de l'estimation de la *loi conditionnelle* de Y sachant X. Ceci est le sujet des résultats exposés dans Chardon *et al.* (2024), où les auteurs obtiennent des vitesses optimales pour l'excès de risque logistique de l'estimateur du maximum de vraisemblance. Ce dernier est le minimiseur du risque associé à la perte logistique. Notre motivation ici est entre autres de traduire ces bornes optimales en estimation de densité en des vitesses optimales pour la classification binaire par règle "plug-in".

Dans ce texte, nous nous intéressons d'abord au modèle logistique et nous nous concentrons sur le cas d'un design gaussien. La loi du couple (X, Y) est alors entièrement caractérisée par le paramètre inconnu θ^* tel que $\eta(X) := \mathbb{P}(Y = 1|X) = \sigma(\langle \theta^*, X \rangle)$. Nous montrons que dans ces conditions, l'excès de risque (en classification) se comporte comme l'erreur quadratique d'estimation de la direction du paramètre θ^*.

Ensuite, nous généralisons ce résultats aux cas où X n'est pas nécessairement gaussien (et $\mathcal{X}$ n'est pas nécessairement euclidien) et où la fonction de régression est de la forme $\eta(X) = g(f^*(X))$, avec g est une fonction lipschitzienne vérifiant $g(0) = 1/2$. Nous montrons que sous certaines hypothèses sur X et une hypothèse de marge, l'excès de risque en classification est à nouveau dominé par l'erreur d'estimation de la "direction" de f^*. Ce résultat doit être mis en perspective du lemme de Zhang.

2 Borne optimale pour le modèle logistique gaussien

Dans cette section nous nous restreignons au cas du modèle logistique gaussien. Plus spécifiquement, l'espace des variables explicatives est $\mathcal{X} = \mathbb{R}^d$. On suppose le design gaussien (sans perte de généralité $X \sim \mathsf{N}(0, I_d)$), et la fonction de regression donnée par $\eta(X) = \sigma(\langle \theta^*, X \rangle)$. On notera simplement $L(u)$ pour le risque du classifieur linéaire associé à $u \in S^{d-1}$, c'est-à-dire $L(u) = \mathbb{P}(Y\langle u, X \rangle < 0)$. On note $u^* = \theta^*/\|\theta^*\|$ la direction du paramètre inconnu.

Proposition 1. *Il existe une constante absolue C telle que pour tout $u \in S^{d-1}$ tel que $\|u - u^*\| \leqslant 1$,*

$$L(u) - L(u^*) \leqslant C\|\theta^*\| \cdot \|u - u^*\|^2 . \tag{2.1}$$

Ceci montre entre autres choses que l'estimateur par maximum de vraisemblance (EMV) en régression logistique atteint la meilleure vitesse possible parmi les classifieurs linéaires. En effet, le Théorème 1 dans Chardon *et al.* (2024) montre que dès que l'EMV existe, il satisfait une borne d'excès de risque (logistique) $O((d + \log \delta^{-1})/n)$ avec probabilité $1 - \delta$.

De manière équivalente (voir la discussion après le Théorème 1 dans la référence), la direction de l'EMV $u^* = \theta^*/\|\theta^*\|$ est correctement estimée, au sens où avec probabilité $1-\delta$,

$$\|\widehat{u}_n - u^*\| \leqslant C' \sqrt{\frac{d + \log(1/\delta)}{Bn}} \,. \tag{2.2}$$

Finalement, il s'en suit que sur l'évènement où l'EMV existe, il est localisé et vérifie

$$L(u) - L(u^*) \lesssim \|\theta^*\| \frac{d + \log(1/\delta)}{Bn} \lesssim \frac{d + \log(1/\delta)}{n} \,.$$

Notons qu'avec une application directe du lemme de Zhang pour la perte logistique, on obtiendrait une borne en $O(\sqrt{d/n})$, ce qui est la vitesse pessimiste attendue en classification.

La preuve de la Proposition 1 repose sur une formule classique de représentation de l'excès de risque sur laquelle repose également la preuve du lemme de Zhang.

Lemme 1. *Pour tout prédicteur $f : \mathbb{R}^d \to \{-1, 1\}$,*

$$L(f) - L(f^*) = \mathbb{E}\big[|2\eta(X) - 1| \mathbf{1}(f(X) \neq f^*(X))\big] \,.$$

Une analyse fine de cette représentation conduit au résultat de la Proposition 1. Cette formule pour l'excès de risque étant très géénérale, elle permet de comprendre les propriétés essentielles du design P_X et de la fonction de régression η qui permettent de généraliser à d'autres cadres.

3 Nouvelles formes du lemme de Zhang

Dans cette section nous étendons les résultats précédents à des modèles autres que le logistique et pour des designs plus généraux. L'hypothèse clé est celle dite de marge, introduite par Mammen et Tsybakov (1999) et réutilisée par Tsybakov (2004) ou Audibert et Tsybakov (2007) pour l'obtention de vitesses rapides en classification. Nous supposons ici que $\eta(X) = g(f^*(X))$, où g est une fonction ρ-lipschitzienne vérifiant $g(0) = 1/2$. On suppose de plus qu'il existe $\beta^*, M, \gamma > 0$ tels que pour tout $t > 0$,

$$\mathbb{P}(|f^*(X)/\beta^*| \leqslant t) \leqslant M t^\gamma \,. \tag{3.1}$$

Sous cette hypothèse, nous obtenons le résultat suivant.

Proposition 2. *Sous les hypothèses précédentes, tout prédicteur f pour lequel il existe $\beta, K > 0$ tels que*

$$\Big\| \frac{f(X)}{\beta} - \frac{f^*(X)}{\beta^*} \Big\|_{\psi_\alpha} \leqslant K \,,$$

a un excès de risque satisfaisant

$$L(f) - L(f^*) \leqslant C_\alpha \rho \beta^* K^{1+\gamma} \max\Big\{1, M(\gamma \log(1/K))^{2/\alpha}\Big\} \,, \tag{3.2}$$

où C_α dépend exclusivement de α.

Ce résultat s'applique au modèle logistique avec design non gaussien. Notons que la condition (3.1) est la forme standard de la condition de marge et est légèrement plus forte que l'hypothèse dite de marge unidimensionnelle dans Chardon *et al.* (2024) qui apparait dans la définition des designs dits "réguliers". Comme corollaire de la Proposition 2, nous obtenons que dans le modèle logistique bien spécifié avec design régulier (condition de marge classique (3.1)), avec probabilité au moins $1 - \exp(-d/c)$,

$$L(\widehat{\theta}_n) - L(\theta^*) \lesssim \text{polylog}(B)\, \frac{d}{n} \log(n/d)\,, \tag{3.3}$$

à des constantes près, ne dépendant que du design. Similairement à la section précédente, nous obtenons un gain significatif dans la vitesse obtenue, puisque le lemme de Zhang aurait à nouveau donné une vitesse en $O(\sqrt{d/n})$.

Bibliographie

Audibert, J.-Y. et Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633. 4

Bartlett, P. L., Jordan, M. I. et McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156. 2

Chardon, H., Lerasle, M. et Mourtada, J. (2024). Finite-sample performance of the maximum likelihood estimator in logistic regression. *arXiv preprint arXiv: 2411.02137*. 3, 5

Mammen, E. et Tsybakov, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829. 4

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166. 4

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. Springer. 2

Vapnik, V. N. et Chervonenkis, A. Y. (1974). *Theory of Pattern Recognition: Statistical Learning Problems (in Russian)*. Nauka, Moscow. 2

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85. 1, 2

Set To Be Fair: Set-Valued Classification Under Fairness Constraints

Eyal Cohen [1] & Christophe Denis [2] & Mohamed Hebiri [3]

[1] *LPSM, France, eyal.cohen@lpsm.paris*
[2] *SAMM, France, christophe.denis1@univ-paris1.fr*
[3] *LAMA, France, mohamed.hebiri@univ-eiffel.fr*

Résumé. La classification ensembliste est utilisée dans un cadre de classification multi-classe où des confusions peuvent survenir et induire en erreur. Cet article aborde le problème de la classification ensembliste sous des contraintes d'équité et de taille moyenne. Nous considérons la contrainte dite de parité démographique, imposant que le classifieur ensembliste assure une représentation équitable des classes entre les différentes valeurs de l'attribut sensible. Nous établissons d'abord une forme explicite du classifieur optimal. À partir de cette expression, nous proposons une procédure empirique basée sur une approche plug-in pour construire un prédicteur empirique. Nous établissons des vitesses de convergence sur l'excès de risque ainsi que sur les violations des contraintes de taille et d'équité. Enfin, des expériences empiriques mettent en avant les performances de l'algorithme proposé.

Mots-clés. Machine Learning, Statistiques, Classification, Classification ensembliste

Abstract. Set-valued classification is used in a multiclass classification setting where confusion can happen and be misleading. This paper tackles the problem of set-valued classification under fairness and expected size constraints. We consider the demographic parity constraint, where the set-valued classifier must achieve an equal representation of each classes between the values of the sensitive attribute. We first provide a closed form expression of the optimal fair set-valued classifier. Based on this expression, we provide a data-driven procedure relying on a plug-in approach to build an empirical predictor. Notably, we establish rates of convergence on the excess risk and on both size and fairness constraint violations. Finally, empirical experiments highlight the performances of the proposed algorithm.

Keywords. Machine Learning, Statistics, Classification, Set-Valued Classification

1 Introduction

Avec l'essor des modèles d'apprentissage automatique dans des domaines sensibles comme la santé, la justice et la finance, l'objectif de réduire les biais et discriminations est une question majeure dans la communauté statistique. En classification automatique, garantir l'équité implique souvent d'ajouter des contraintes qui peuvent dégrader les performances du modèle (Denis, Elie, et al. 2023). Dans ce travail, nous explorons une approche de classification ensembliste (Denis et Hebiri 2017) équitable, où le classifieur peut retourner un ensemble

de classes potentielles, ce qui permet de réduire le risque de mauvaise classification. Cette approche offre une alternative à l'équité approximative, pour la réduction de l'impact de l'équité sur les performances du classifieur.

Nous introduisons une version du critère d'équité de parité démographique adaptée à ce cadre et nous déterminons le classifieur optimal respectant simultanément cette contrainte et une contrainte sur la taille moyenne des ensembles prédits. Nous proposons ensuite une procédure empirique permettant d'estimer ce classifieur à partir des données. Enfin, nous analysons les performances de notre méthode sur des données simulées et réelles, mettant en évidence que l'équité peut être atteinte sans dégrader significativement les performances du classifieur.

2 Modèle et méthodologie

2.1 Formulation du problème

Soit $(X, S, Y) \in \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$ un triplet aléatoire où X sont les covariables, S est l'attribut sensible (par exemple, genre ou origine ethnique) et Y est la variable à prédire. En classification ensembliste, un classifieur $\Gamma : \mathcal{X} \times \mathcal{S} \to 2^{\mathcal{Y}}$ associe à tout couple (x, s) un ensemble $\Gamma(x, s)$ de classes possibles. On note $\mathbf{\Gamma}$ l'ensemble des classifieurs ensemblistes.

Dans ce contexte nous cherchons un classifieur ensembliste équitable qui soit optimal relativement au risque de mauvaise classification sous contrainte de contrôle de sa taille moyenne. Ceci motive les deux définitions suivantes.

Définition 2.1. Soit $\Gamma \in \mathbf{\Gamma}$ un classifieur ensembliste. On dit que Γ respecte la contrainte de taille moyenne β si:

$$\mathbb{E}_{X,S}\left[|\Gamma(X, S)|\right] \leq \beta.$$

Définition 2.2. Soit $\Gamma \in \mathbf{\Gamma}$ un classifieur ensembliste. On dit que Γ est équitable pour la contrainte de parité démographique si:

$$\forall k \in [K], \forall s \in \mathcal{S} : \mathbb{P}_{X|S=s}\left(k \in \Gamma(X, s)\right) - \mathbb{P}_{X,S}\left(k \in \Gamma(X, S)\right) = 0.$$

Le classifieur optimal, Γ^*, s'écrit alors:

$$\arg\min \Big\{ \mathbb{P}\left(Y \notin \Gamma(X, S)\right) : \Gamma \in \mathbf{\Gamma} \text{ tq.} \mathbb{E}_X\left[|\Gamma(X, S)|\right] \leq \beta \text{ et } \forall k \in [K], \forall s \in \mathcal{S}, \\ \mathbb{P}_{X|S=s}\left(k \in \Gamma(X, s)\right) - \mathbb{P}_{X,S}\left(k \in \Gamma(X, S)\right) = 0 \Big\}. \quad (1)$$

2.2 Classifieur optimal

L'objectif de cette section est d'établir une expression explicite du prédicteur optimal Γ^*. Soit $(\pi_s)_{s\in\mathcal{S}}$ la loi discrète de l'attribut sensible S. Soient les fonctions caractérisées par:

$$\begin{aligned}\forall t \in [0,1], \quad \overline{F}_k(t) &= \mathbb{P}_{X,S}\left(p_k(X,S) \geq t\right), \\ \overline{F}_{k,s}(t) &= \mathbb{P}_{X|S=s}\left(p_k(X,s) \geq t\right), \\ G(t) &= \sum_{k=1}^{K} \overline{F}_k(t).\end{aligned}$$

Nous introduisons l'hypothèse suivante:

Hypothèse 1. Les variables aléatoires $p_k(X,S)$ admettent une densité strictement positive.

Sous l'hypothèse 1, nous obtenons une forme explicite du prédicteur optimal Γ^*.

Théorème 2.3. *Soient $\lambda \in \mathbb{R}_+$ et $\alpha_{k,s} \in \mathbb{R}$ pour $k \in [K]$ et $s \in \mathcal{S}$ les paramètres du lagrangien associé à l'Equation (1) que l'on annote d'une * quand ils maximisent ce dernier. Alors, sous l'hypothèse 1, le classifieur ensembliste optimal sous contraintes de parité démographique est caractérisé par: pour tout $k \in [K], s \in \mathcal{S}$,*

$$\Gamma^*_{\lambda^*,\gamma^*}(x,s) = \left\{ k \in [K] : p_k(x,s) \geq \lambda^* + \frac{\gamma^*_{k,s}}{\pi_s} \right\},$$

avec $p_k(x,s) = \mathbb{P}\left(Y = k | X = x, S = s\right)$ et $\gamma^*_{k,s} = \alpha^*_{k,s} - \sum_{s\in\mathcal{S}} \alpha^*_{k,s}$. Les paramètres λ^*, γ^* sont uniquement définis comme suit:

$$\lambda^* = G^{-1}(\beta), \quad \gamma^*_{k,s} = \pi_s \left(\overline{F}^{-1}_{k,s}\left(\overline{F}_k(\lambda^*)\right) - \lambda^* \right).$$

Comme dans (Denis et Hebiri 2017), le classifieur optimal sous contrainte de parité démographique se traduit par un seuillage pour chaque couple (k,s), là où le classifieur optimal sans cette contrainte était caractérisé par un seuil commun. De ce point de vue, la correction d'équité s'exprime simplement à travers l'ajustement des seuils.

3 Procédure empirique

Au vu du résultat du Theorem 2.3, nous proposons une approche de type plug-in pour l'estimation du classifieur. Nous utilisons pour cela (a) un échantillon labelisé D_n permettant de construire les estimateurs $\widehat{p}_k$ des fonctions p_k; (b) un échantillon non-labelisé D_M pour estimer les paramètres

$$\widehat{\delta}_{k,s} = \widehat{\overline{F}}^{-1}_{k,s}\left(\widehat{\overline{F}}_k\left(\widehat{G}^{-1}(\beta)\right)\right).$$

Soient M_s le nombre d'éléments de D_M tels que $S = s$ et $\widehat{\pi}_s$, la proportion empirique de cet événement. Nous définissons les fonctions empiriques suivantes:

$$\forall (k,s) \in [K] \times \mathcal{S}, \forall t \in [0,1], \quad \widehat{\overline{F}}_{k,s}(t) = \frac{1}{M_s} \sum_{i=1}^{M} \mathbb{1}_{S_i = s} \mathbb{1}_{\widehat{p}_k(X_i,S_i) \geq t},$$

$$\widehat{\overline{F}}_{k}(t) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}_{\widehat{p}_k(X_i,S_i) \geq t},$$

$$\widehat{G}(t) = \sum_{k=1}^{K} \widehat{\overline{F}}_{k}(t),$$

et les inverses généralisées:

$$\forall (k,s) \in [K] \times \mathcal{S}, \forall u \in \mathbb{R}, \quad \widehat{\overline{F}}^{-1}_{k,s}(u) = \inf \left\{ t \in [0,1] : \widehat{\overline{F}}_{k,s}(t) \geq u \right\}$$

$$\widehat{G}^{-1}(u) = \inf \left\{ t \in [0,1] : \widehat{G}(t) \geq u \right\}.$$

L'estimateur résultant offre des performances théoriques intéressantes. Nous commençons par borner les violations de contraintes. Soit $\widehat{\Gamma}$ l'estimateur de Γ^* construit avec la procédure empirique décrite ci-dessus, de telle sorte que

$$\widehat{\Gamma}(x,s) = \left\{ k \in [K] : \widehat{p}_k(x,s) \geq \widehat{\delta}_{k,s} \right\}.$$

Proposition 3.1. Pour tout $\beta \in [0, K]$:

$$\mathbb{E}_{X,S} \left[\left| \widehat{\Gamma}(X,S) \right| \right] = \beta + \mathcal{O}\left(\frac{K}{\sqrt{M}} \right).$$

Proposition 3.2. Soit, pour tout classifieur $\Gamma \in \mathbf{\Gamma}$:

$$U(\Gamma) = \max_{\substack{k \in [K] \\ (s,s') \in \mathcal{S}^2}} \left| \mathbb{P}_{X|S=s} \left(k \in \Gamma(X,s) \right) - \mathbb{P}_{X|S=s'} \left(k \in \Gamma(X,s') \right) \right|$$

sa mesure d'inéquité. On a:

$$\mathbb{E}_{X,S} \left[U(\widehat{\Gamma}) \right] = \mathcal{O}\left(\frac{K}{\sqrt{M}} \right).$$

Ces propositions indiquent que les violations de contrainte ne dépendent pas des estimateurs des fonctions p_k, mais seulement du nombre de labels et de la taille de l'échantillon d'entraînement, D_M.

Pour le borner le risque, nous introduisons une seconde hypothèse:

Hypothèse 2. La distribution $\mathbb{P}_{X|S=s}$ satisfait l'hypothèse de marge de niveau τ, avec $\tau > 0$, pour les variables aléatoires $p_k(X,s)$ relativement à $G^{-1}_{k,s}(\beta)$. C'est à dire qu'il existe $C > 0$ et $t_0 \in (0,1)$ tels que, pour tout $t \leq t_0$:

$$\mathbb{P}_{X|S=s} \left(\left| p_k(X,s) - G^{-1}_{k,s}(\beta) \right| \leq t \right) \leq C t^{\tau},$$

où $G^{-1}_{k,s}(\beta) = \overline{F}^{-1}_{\widehat{p}_k|S=s} \left(\overline{F}_{\widehat{p}_k} \left(G^{-1}(\beta) \right) \right)$

Théorème 3.3. *Soit le risque pénalisé caractérisé par: pour tout* $\Gamma \in \mathbf{\Gamma}$,

$$\mathcal{R}^*(\Gamma) = \mathcal{R}(\Gamma) + \lambda^* \mathbb{E}_{X,S}\left[|\Gamma(X,S)|\right] + \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \alpha^*_{k,s}\left(\mathbb{P}_{X|S=s}\left(k \in \Gamma\right) - \mathbb{P}_{X,S}\left(k \in \Gamma\right)\right).$$

L'estimateur $\widehat{\Gamma}$ *vérifie:*

$$\mathbb{E}\left[\mathcal{R}^*(\widehat{\Gamma})\right] - \mathcal{R}(\Gamma^*) \lesssim K\left(\max_{s\in\mathcal{S}} ||\widehat{p} - p||^{\tau}_{\infty,\mathbb{P}_{X|S=s}} + G^{-1}(\beta)\frac{K|\mathcal{S}|}{\sqrt{M}} + \frac{1}{\sqrt{M}}\right).$$

4 Résultats

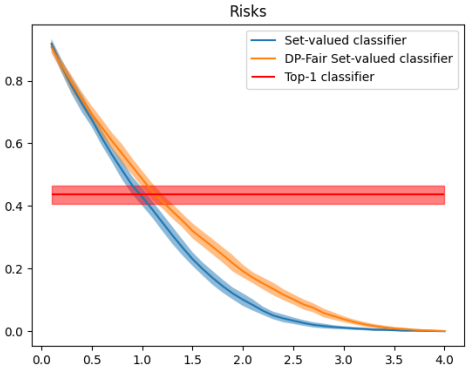
Nous présentons ici les résultats obtenus par la procédure décrite en Section 3. Nous utilisons le jeu de données DRUG concernant la consommation de drogue en fonction de mesures de personnalités (impulsivité, neuroticisme, ...), l'âge, le genre, l'ethnie, etc. sur lequel nous appliquons un pré-processing réduisant notamment le nombre des possibilités de consommations de cannabis de 7 à 4 (jamais consommé, consommé il y a plus d'un an, consommé il y a moins d'un an, consommé dans les dernière 24h) et essayons de prédire cette consommation. Nous prenons dans la figure 1 l'âge comme attribut sensible et observons que le prix en terme de risque est faible pour une nette amélioration de l'équité. De plus, la contrainte de taille n'est pas impactée par la correction d'équité, ce qui est cohérent avec le résultat de la Proposition 3.1 qui nous donne la même vitesse que pour le classifieur sans contrainte de parité (Denis et Hebiri 2017, proposition 5).

Conclusion

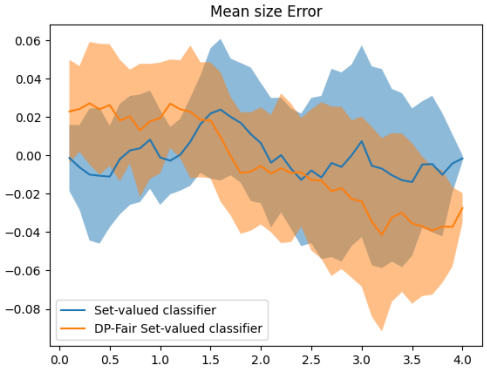
Dans ce travail, nous proposons une approche de classification ensembliste sous une contrainte de parité démographique. Nous avons obtenu une forme explicite du classifieur optimal et des vitesses de convergence sur le risque et les contraintes. Les résultats empiriques montrent que l'équité peut être nettement améliorée sans dégrader significativement les performances du classifieur. Les perspectives de ce travail sont nombreuses, notamment l'extension de cette approche à d'autres contraintes d'équité, ou l'approche de minimisation de la taille pour un risque contrôlé.

References

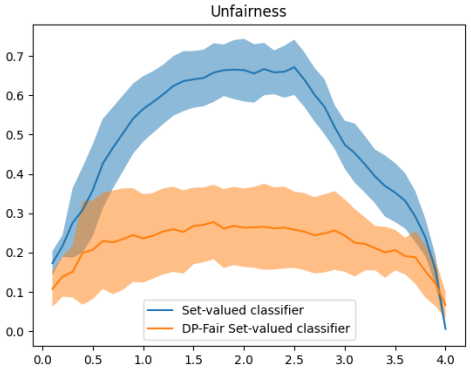
Denis, Christophe, Romuald Elie, et al. (Mar. 2023). "Fairness guarantee in multi-class classification". In: arXiv:2109.13642. arXiv:2109.13642 [math, stat]. DOI: 10.48550/arXiv.2109.13642. URL: http://arxiv.org/abs/2109.13642.



(a) Risque de mauvaise classification



(b) Erreur de taille moyenne



(c) Inéquité

Figure 1: Comparaison du classifieur ensembliste avant et après correction de l'équité sur l'âge, séparé par métrique.

Denis, Christophe et Mohamed Hebiri (Dec. 2017). "Confidence sets with expected sizes for Multiclass Classification". en. In: arXiv:1608.08783. arXiv:1608.08783 [math, stat]. URL: http://arxiv.org/abs/1608.08783.

Aggrégation d'ensemble de prédiction corrélés par la classe

Jean-Baptiste Fermanian [1] & Mohamed Hebiri[2] & Joseph Salmon[3]

[1] *IMAG, Univ Montpellier, CNRS, Inria, Montpellier, France, jean-baptiste.fermanian@inria.fr*
[2] *LAMA, Université Gustave Eiffel, France, mohamed.hebiri@univ-eiffel.fr*
[3] *IMAG, Univ Montpellier, CNRS, Inria, Montpellier, France, joseph.salmon@inria.fr*

Résumé. Nous construisons une méthode d'agrégation pour classifier un objet pour lequel nous avons à disposition plusieurs observations et une prédiction pour chacune d'entre elles. Notre approche repose sur la quantification de l'incertitude de chacune de ces prédictions à l'aide de p-valeurs conformelles. Nous proposons plusieurs méthodes d'agrégation basée sur l'étude de la loi jointe du vecteur de ces p-valeurs à partir des travaux de Marques F (2024) et Gazin et al. (2023). Nous en déduisons un ensemble de classes satisfaisant des garanties de couverture conditionnelle. Notre approche a pour but d'être utilisée pour les données de PlantNet où chacune des observations est une photographie d'une même plante que l'on cherche à classifier tout en fournissant une quantification de l'erreur de la prédiction.

Mots-clés. Prédiction conforme, classification multi-classe, agrégation, p-valeur conformelle, PlantNet.

Abstract. We propose an aggregation method to classify an object for which we have several observations and a prediction for each of them. Our approach is based on quantifying the uncertainty of each of these predictions using conformal p-values. We propose several aggregation methods based on the study of the joint distribution of the vector of these p-values, based on the works of Marques F (2024) and Gazin et al. (2023). We derive a set of possible labels that satisfies conditional coverage guarantees. Our approach is intended for use with PlantNet data, where each observation is a picture of the same plant that we seek to classify, while providing a quantification of the prediction error.

Keywords. Conformal prediction, multilabel classification, aggregation, conformal p-value, PlantNet.

1 Présentation du problème

Nous considérons le problème de classification multi-classe. Nous avons à disposition un classifieur $\mathcal{A} : \mathcal{X} \to [0,1]^K$ qui associe à une image, une probabilité d'être dans une des classes. Ce classifieur est par exemple un réseau de neurones dont la dernière couche est un softmax. Nos données consistent en un échantillon de *calibration* $(X_i, Y_i)_{i=1}^n$ de distribution $\mathbb{P} = \mathbb{P}_Y \mathbb{P}_{X|Y}$ et des observations $(X_{n+j})_{j=1}^m$ partageant la même classe Y_{n+1}. Formellement,

ces points sont tirés selon la loi $\mathbb{P}_{X|Y=Y_{n+1}}$. Notre objectif est de construire, pour $\alpha \in]0,1[$, un ensemble de prédiction $\mathcal{C}_\alpha = \mathcal{C}_\alpha\big((X_i,Y_i)_{i=1}^n, (X_{n+j})_{j=1}^m\big) \subset [\![K]\!] \triangleq \{1,\dots,K\}$, tel que :

$$\mathbb{P}[Y_{n+1} \in \mathcal{C}_\alpha] \geq 1-\alpha. \tag{1}$$

Cette condition est connue sous le nom de *couverture marginale* et est un objectif classique des approches de prédiction conforme (Papadopoulos et al., 2002 ; Vovk et al., 2005). Lorsque les classes sont très déséquilibrées, pour ne pas pénaliser les classes de petite taille, nous demanderons à notre méthode de satisfaire une couverture conditionnelle, c'est-à-dire pour toute classe $y \in [\![K]\!]$:

$$\mathbb{P}[y \in \mathcal{C}_\alpha | Y_{n+1} = y] \geq 1-\alpha. \tag{2}$$

Cette condition est particulièrement nécessaire pour les données de PlantNet où les classes observées sont très déséquilibrées (Garcin et al., 2021). Nous demandons alors d'avoir une garantie de couverture pour toutes les classes quelque soit leur probabilité d'apparition.

2 Prédiction conforme en classification

Les méthodes classiques de prédiction conforme reposent sur l'évaluation d'une fonction de *scores* sur l'échantillon de calibration. Dans le cadre de classification multiclasse, en supposant que l'output de l'algorithme est un vecteur de probabilité sur l'ensemble des classes, nous pouvons considérer plusieurs fonctions de scores s :

— *Threshold :* $s(x,y) = \mathcal{A}_y(x)$.
— *Adaptive prediction sets (APS) :* $s(x,y) = \sum_{y'} \mathcal{A}_{y'}(x)\mathbf{1}\{\mathcal{A}_{y'}(x) \leq \mathcal{A}_y(x)\}$.
— *Rank-based (RANK) :* $s(x,y) = \sum_{y'} \mathbf{1}\{\mathcal{A}_{y'}(x) < \mathcal{A}_y(x)\}$.

Notons $S_i = s(X_i,Y_i)$ pour $1 \leq i \leq n$ la fonction de score évaluée sur l'échantillon de calibration. L'ensemble de prédiction $\mathcal{C}_\alpha(x) = \big\{y : s(x,y) \geq S_{(\lfloor (n+1)\alpha \rfloor)}\big\}$, où $S_{(k)}$ désigne le k-ième plus grand score, est *marginalement* valide (i.e. satisfait (1)). Une manière simple d'obtenir un ensemble de prédiction conditionnellement valide est de considérer les scores par classe. Notons n_y le nombre de points de la classe y dans l'échantillon de calibration et S_i^y pour $i \in [\![n_y]\!]$ les scores de classe y. L'ensemble

$$\mathcal{C}_\alpha^c = \Big\{y : s(x,y) \geq S^y_{(\lfloor (n_y+1)\alpha \rfloor)}\Big\}$$

est *conditionnellement* valide, au sens de l'Équation (2). Pour ces deux ensembles, seule une hypothèse d'échangeabilité sur les données est nécessaire.

Ces ensembles peuvent se réécrire en terme de p-valeurs conformelles. Nous les définirons ici, pour $y \in [\![K]\!]$ et $j \in [\![m]\!]$, par :

$$p_j(y) = \frac{1}{n_y}\sum_{i=1}^{n_y} \mathbf{1}\{s(X_{n+j},y) \geq S_i^y\}\,. \tag{3}$$

Nous allons considérer deux types d'approches pour agréger les ensembles de prédictions des différentes observations. Les premières se relient aux approches proposées par Gasparin

et Ramdas (2024) et utilisent les statistiques de comptage $S(y) = \sum_{j=1}^{m} \mathbf{1}\{y \in \mathcal{C}_\alpha^c(X_{n+j})\}$ comptant le nombre d'ensembles de prédictions contenant la classe y. Nos autres méthodes vont utiliser le vecteur ordonné des p valeurs $p(y) = (p_{(1)}(y), \ldots, p_{(m)}(y))$ dont la distribution est connue et indépendante de la distribution originelle des données sous certaines hypothèses. Ces méthodes reposent sur des résultats de Gazin et al. (2023).

3 Vote par majorité

Gasparin et Ramdas (2024) agrègent des ensembles de prédiction de niveau α de dépendance arbitraire à partir des statistiques $S(y)$. L'idée naturelle est de compter le nombre d'ensemble acceptant une classe y et le conserver si ce nombre dépasse un certain seuil, la moitié dans le cas le plus simple. Dans notre contexte, ce seuil peut être augmenté, ce qui réduit les ensembles, tout en gardant une couverture valide. Si les scores sont échangeables par classe, i.e. si

$$\text{le vecteur } \left(X_1^{Y_{n+1}}, \ldots, X_{n_{Y_{n+1}}}^{Y_{n+1}}, X_{n+1}, \ldots, X_{n+m}\right) \text{ est échangeable,} \tag{4}$$

et que les scores n'aient pas d'égalité, alors sachant $Y_{n+1} = y$, la statistique $S(y)$ suit une distribution Béta-Binomiale $\text{BetaBin}(m, n_y + 1 - k_y, k_y)$ où $k_y = \lfloor (n_y+1)\alpha \rfloor$), (voir par exemple Marques F, 2024 ; Vovk, 2012). Lorsqu'il y a des égalités, la statisque $S(y)$ domine cette distribution. Pour améliorer les méthodes proposées par Gasparin et Ramdas (2024), nous utilisons comme seuil les quantiles des lois Béta-Binomiale. Ces lois peuvent être approchées par des lois Binomiale lorsque le nombre de points de calibration est important.

Proposition 3.1. *Supposons* (4). *Soit* $\alpha \in (0,1)$ *et* $\mathcal{C}^{BB} = \{y : S(y) \geq q_\alpha(y)\}$ *où* $q_\alpha(y)$ *le quantile* α *de la loi BetaBin*$(m, n_y + 1 - k_y, k_y)$, *alors, pour tout* $y \in [\![K]\!]$ *:*

$$\mathbb{P}\big[y \in \mathcal{C}^{BB} | Y_{n+1} = y\big] \geq 1 - \alpha\,.$$

De plus, si q_α *est le quantile* α *de la loi Binomiale* $\mathcal{B}(m, 1-\alpha)$ *et* $\mathcal{C}^{Bin} = \{y : S(y) \geq q_\alpha\}$, *alors, pour tout* $y \in [\![K]\!]$ *:*

$$\mathbb{P}\big[y \in \mathcal{C}^{BB} | Y_{n+1} = y\big] \geq 1 - \alpha - \frac{m}{n_y}\,.$$

Ces approches simples sont des premiers points de références pour les approches suivantes.

4 Approche p-valeurs

Les méthodes précédentes utilisent les p valeurs conformelles pour en déduire des intervalles de prédictions pour chaque image et ensuite les agréger à l'aide des statistiques de comptage $S(y)$. Les approches suivantes vont directement utiliser les vecteurs des p valeurs ordonnées $p(y)$ pour construire un ensemble de prédiction conditionnellement valide.

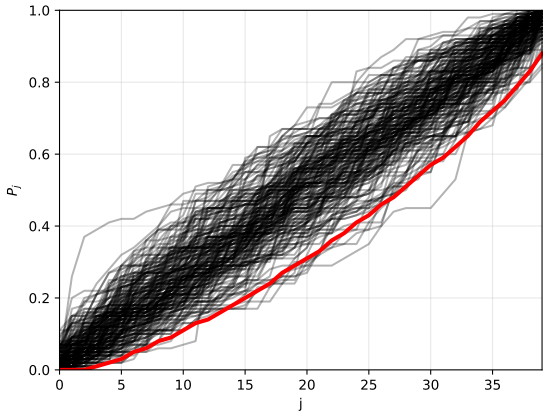


FIGURE 1 – 200 réalisations de $P \sim \mathcal{U}(A_{n,m})$ (en noir), et, en rouge, $E_\alpha^{n,m}$ estimée par Monte-Carlo pour $m = 40$ et $n = 100$.

4.1 Loi exacte et randomisation

Si les scores sur la calibration et les observations sont presque sûrement différents, à partir des travaux de GAZIN et al. (2023), nous démontrons que conditionnellement à la classe Y_{n+1} le vecteur des p-valeurs ordonnées $p(Y_{n+1})$ suit une loi uniforme sur l'ensemble des vecteurs ordonnées $A_{n_{Y_{n+1}},m}$ défini par :

$$A_{n,m} = \left\{ r \in \left\{ \frac{i}{n} : 0 \leq i \leq n \right\}^m : r_1 \leq \ldots \leq r_m \right\}.$$

En Figure 1, nous traçons plusieures réalisations de la loi uniforme sur $A_{n,m}$. En cas d'égalité de certains des scores (par exemple avec la fonction de scores de rang), on peut utiliser la randomisation $p^{rd}(y)$ définie pour $j \in [\![m]\!]$ et $y \in [\![K]\!]$ par :

$$p_j^{rd}(y) = \frac{1}{n_y} \sum_{i=1}^{n_y} \mathbf{1}\{s(X_{n+j}, y) > S_i^y\} + \mathbf{1}\{s(X_{n+j}, y) = S_i^y\} \mathbf{1}\{U_{n+j} \geq U_i\},$$

où U_i est une suite de variables aléatoires i.i.d. de loi uniforme sur $[0, 1]$. Soit $p^{rd}(y)$ le vecteur ordonné, alors $p^{rd}(Y_{n+1}) \sim \mathcal{U}(A_{n_{Y_{n+1}},m})$ et le vecteur est dominé par le vecteur $p(y)$, i.e. pour tout $j \in [\![K]\!]$, $p_j(y) \geq p_j^{rd}(y)$. Nous construisons plusieurs méthodes d'agrégation utilisant l'un ou l'autre de ces deux objets.

4.2 Approche quantile

Soit $\succcurlyeq$ un ordre sur les vecteurs de $A_{n,m}$. Nous allons chercher à construire une enveloppe $E_\alpha^{n,m} \in A_{n,m}$, tel que pour $P \sim \mathcal{U}(A_{n_y,m})$, on ait $\mathbb{P}[P \succcurlyeq E_\alpha^{n,m}] \geq 1 - \alpha$. La trajectoire $E_\alpha^{n,m}$ peut être vue comme un quantile multidimensionel pour l'ordre $\succcurlyeq$. L'ensemble de prédiction résultant est alors $\mathcal{C}_\alpha^{\succcurlyeq} = \left\{ y : p(y) \succcurlyeq E_\alpha^{n_y,m} \right\}$.

L'ordre $\succcurlyeq$ que nous considérons est la domination d'une courbe par une autre, i.e. pour $p, q \in A_{n,m}$, $p \succcurlyeq q$ si et seulement si $p_j \geq q_j$ pour tout $j \in [\![m]\!]$. La trajectoire $E_\alpha^{n,m}$ est estimée par une méthode de Monte Carlo similaire à celle utilisée par Fermanian et al. (2025). Pour un telle enveloppe, on obtient la garantie de coverage conditionnelle :

$$\mathbb{P}\big[p(y) \succcurlyeq E_\alpha^{n_y,m} | Y_{n+1} = y\big] \geq \mathbb{P}\big[p^{rd}(y) \succcurlyeq E_\alpha^{n_y,m} | Y_{n+1} = y\big] \geq 1 - \alpha\,.$$

4.3 Approche score

Comme les lois $\mathcal{U}(A_{n_y,m})$ sont connues et facilement simulables, il est possible d'utiliser à nouveau une méthode de prédiction conforme sur les p-valeurs ordonnées. Soit $S : [0,1]^m \to \mathbb{R}$ une fonction de score pour les vecteurs de p-valeurs, on génère pour chaque classe $y \in [\![K]\!]$ un échantillon $(P_i^y)_{i=1}^N$ de taille N et de distribution uniforme sur $A_{n_y,m}$. On utilise ces échantillons comme des échantillons de calibration. Ainsi l'ensemble

$$\mathcal{C}^{sc} = \left\{ y : S(p^{rd}(y)) \geq V^y_{\lfloor \alpha(N+1) \rfloor} \right\}$$

est conditionellement valide (Eq. (2)), où $V_i^y = S(P_i^y)$ pour $i \in [\![N]\!]$ et $y \in [\![K]\!]$. Si pour $p, q \in [0,1]^m$, la fonction de score S conserve l'ordre $\succcurlyeq$ (i.e. $p \succcurlyeq q \Rightarrow S(p) \geq S(q)$), on peut alors utiliser le vecteur non randomisé p au lieu de p^{rd} dans $\mathcal{C}^{sc}$. Nous construisons différentes fonctions de scores S sur ces trajectoires considérant différentes notions d'aires ou d'ordre.

5 Expériences sur données synthétiques

Les expériences présentées ici sont sur des données synthétiques. Nous considérons un problème de classification multiclasse avec $K = 10$ classes. Les données sont tirées selon un mélange de gaussiennes isotropes centrées en $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$ choisis aléatoirement. Formellement, pour p une distribution sur $[\![K]\!]$, Y est tiré selon p et $X|Y \sim \mathcal{N}(\mu_Y, I_d)$. Le classifieur $\mathcal{A}$ est un réseau de neurones à 4 couches cachées de largeur 15 entraîné sur 500 données. Le classifieur a un taux d'erreur de 53%.

Les méthodes présentées en Figure 2 sont le vote par majorité (MAJ) de Gasparin et Ramdas (2024), les ensembles $\mathcal{C}^{Bin}$ (Bin) et $\mathcal{C}^{BB}$ (BB), l'approche quantile en utilisant les p-valeurs randomisées (Quant) et non randomisées (QuantRd), l'approche score en utilisant comme score la somme des p-valeurs (L1) et la distance de la trajectoire à l'identité (L1Id). Toutes ces méthodes utilisent le score Threshold.

Nous traçons l'évolution de la longueur moyenne et du coverage en fonction du nombre d'observations m. Les couvertures se maintiennent au dessus de $1 - \alpha = 0.9$ tout en ayant des tailles qui décroient. On observe une amélioration nette par rapport à la méthode naïve de Gasparin et Ramdas (2024) de toutes nos méthodes. Malgré les faibles performances du classifieur, les meilleures ensembles se limitent à 3 classes possibles en moyenne après une dizaines d'observations. Les méthodes utilisant les enveloppes ont de meilleures performances que celles utilisant seulement les statistiques de comptage. Le quantile binomiale approche correctement le quantile de la Beta-Binomiale pour un petit nombre d'observations.

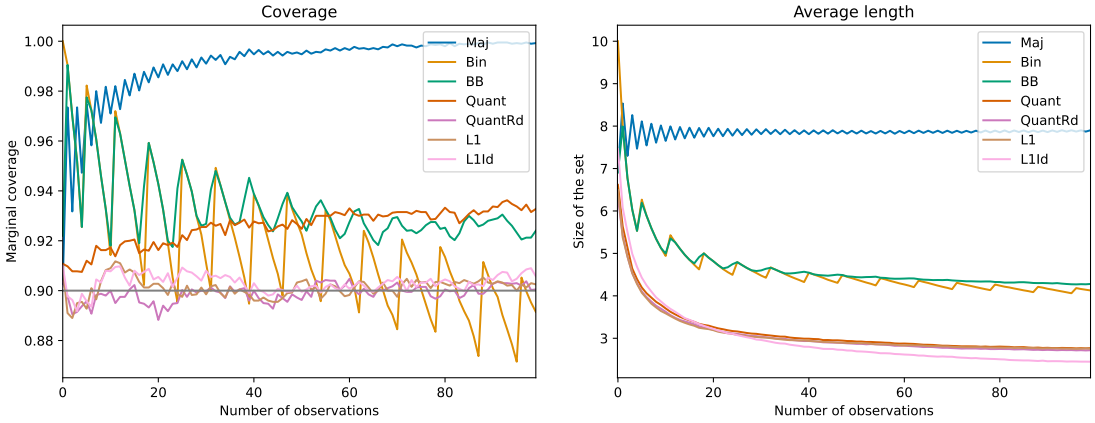


FIGURE 2 – Longueur et couverture marginale en fonction du nombre d'observations pour différentes méthodes évaluées sur 4000 répétitions.

Acknowledgments This work was funded by the French National Research Agency (ANR) through Chaire IA CaMeLOt (ANR-20-CHIA-0001-01).

Références

FERMANIAN, J.-B., HUMBERT, P. et BLANCHARD, G. (2025). Transductive Conformal Inference for Ranking. *arXiv preprint arXiv :2501.11384*.

GARCIN, C., JOLY, A., BONNET, P., AFFOUARD, A., LOMBARDO, J.-C., CHOUET, M., SERVAJEAN, M., LORIEUL, T. et SALMON, J. (2021). Pl@ntNet-300K : a plant image dataset with high label ambiguity and a long-tailed distribution. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

GASPARIN, M. et RAMDAS, A. (2024). Merging uncertainty sets via majority vote. *arXiv preprint arXiv :2401.09379*.

GAZIN, U., BLANCHARD, G. et ROQUAIN, É. (2023). Transductive conformal inference with adaptive scores. *International Conference on Artificial Intelligence and Statistics*.

MARQUES F, P. C. (2024). *Universal distribution of the empirical coverage in split conformal prediction*. arXiv : `2303.02770 [math.ST]`.

PAPADOPOULOS, H., PROEDROU, K., VOVK, V. et GAMMERMAN, A. (2002). Inductive confidence machines for regression. *Machine learning : ECML 2002 : 13th European conference on machine learning Helsinki, Finland, 2002 proceedings 13*. Springer, 345-356.

VOVK, V. (2012). Conditional validity of inductive conformal predictors. *Asian conference on machine learning*. PMLR, 475-490.

VOVK, V., GAMMERMAN, A. et SHAFER, G. (2005). *Algorithmic learning in a random world*. T. 29. Springer.

Clustering

Sélection du nombre de classes pour un algorithme de clustering à larges marges

Félix Laplante [1] & Christophe Ambroise [1]

[1] *Université Paris-Saclay, CNRS, Univ Evry, Laboratoire de Mathématiques et Modélisation d'Évry, France, christophe.ambroise@univ-evry.fr*

Résumé. Ce travail présente Spectral Bridges, un nouvel algorithme de clustering, ainsi que des stratégies dédiées à la sélection du nombre de clusters. Spectral Bridges s'appuie sur les cadres classiques des k-means et du clustering spectral en subdivisant les données en petites régions de Voronoï, qui sont ensuite fusionnées selon une mesure de connectivité. S'inspirant du concept de marge des Machines à Vecteurs de Support (SVM), une approche de clustering non paramétrique est proposée, établissant une marge entre chaque paire de régions de Voronoï. Cette approche permet de délimiter des structures de groupes complexes et non convexes, tout en étant robuste au choix des hyperparamètres.

L'algorithme Spectral Bridges est implémenté en Python (https://pypi.org/project/spectral-bridges) et en R (https://github.com/cambroise/spectral-bridges-Rpackage).

Mots-clés. Clustering spectral, quantification vectorielle, scalable, non paramétrique, sélection de modèle

Abstract. This work introduces Spectral Bridges, a novel clustering algorithm, along with dedicated strategies for selecting the number of clusters. This algorithm builds upon the classical k-means and spectral clustering frameworks by subdividing data into small Voronoï regions, which are subsequently merged according to a connectivity measure. Drawing inspiration from the margin concept in Support Vector Machines (SVMs), a non-parametric clustering approach is proposed, establishing a margin between each pair of Voronoï regions. This approach delineates complex, non-convex cluster structures while remaining robust to hyperparameter selection.

The Spectral Bridges algorithm is implemented in Python (https://pypi.org/project/spectral-bridges) and R (https://github.com/cambroise/spectral-bridges-Rpackage).

Keywords. spectral clustering, vector quantization, scalable, non-parametric, model selection

1 Introduction

Le clustering est une technique clé en analyse exploratoire des données, permettant de regrouper des objets en clusters homogènes. Il est largement utilisé en biologie pour l'analyse

de l'expression génétique (Eisen et al. 1998), en sciences sociales pour la détection de communautés (Latouche et al. 2011), et en psychologie pour identifier des schémas comportementaux. Il est également employé en apprentissage supervisé comme étape de prétraitement pour améliorer la performance et l'interprétabilité des modèles.

Différentes approches de clustering existent, selon la définition de la similarité, que ce soit par des métriques de distance ou des mesures d'affinité. Certaines méthodes sont heuristiques, d'autres optimisent des critères spécifiques et s'appuient sur des modèles statistiques. Le clustering basé sur la densité, comme DBSCAN (Ester et al. 1996), identifie les régions à forte densité. Le clustering basé sur les modèles, tel que les mélanges gaussiens, suppose que les clusters proviennent de distributions de probabilité, estimées par l'algorithme EM (McLachlan and Peel 2000). Les approches géométriques, comme k-means (MacQueen et al. 1967), minimisent la distance intra-cluster, avec une amélioration via k-means++ pour une meilleure initialisation (Arthur and Vassilvitskii 2006). Cependant, k-means suppose des clusters convexes, ce qui peut être contourné par l'utilisation de l'astuce du noyau (Dhillon et al. 2004). Les méthodes basées sur les graphes, telles que le clustering spectral, exploitent les vecteurs propres de la matrice de Laplacien (Shi and Malik 2000), bien que leur application à grande échelle soit coûteuse en calcul (Von Luxburg 2007).

Ce travail introduit une méthode de sélection du nombre de clusters pour l'algorithme Spectral Bridges (Laplante and Ambroise 2024), une approche de clustering spectral conçue pour détecter des clusters non convexes dans de grands jeux de données sans hypothèses paramétriques. L'approche est fondée sur les diagramme de persistance (Chazal et al. 2013) issus de l'analyse de données topologique la (Carlsson 2009). Des comparaisons avec une méthode de type silouhette (Mur et al. 2016) sont présentées.

2 Affinité du pont et clustering spectral

L'algorithme proposé utilise les centroïdes des k-means pour une quantification vectorielle définissant des régions de Voronoï. Une stratégie est alors mise en place pour relier ces régions, en mesurant une affinité basée sur la marge minimale entre paires de classes. Cette affinité est utilisée comme pondération des arêtes d'un graphe complètement connecté, où les sommets sont les régions de Voronoï. Un *clustering spectral* est ensuite appliqué pour obtenir une partition finale.

L'idée de base est de calculer la différence d'inertie entre l'utilisation des centroïdes indépendamment et leur projection sur le segment qui les relie. Si cette différence est faible, cela indique une faible densité entre les classes. À l'inverse, si elle est grande, cela suggère que les deux classes appartiennent à la même région densément peuplée.

Considérons un échantillon $X = (\boldsymbol{x}_i)_{i\in\{1,\cdots,n\}}$ de vecteurs $\boldsymbol{x}_i \in \mathbb{R}^d$ ainsi qu'un ensemble de m vecteurs de codage $(\boldsymbol{\mu}_k)_{k\in\{1,\cdots,m\}}$ définissant une partition $P = \{\mathcal{V}_1, \cdots, \mathcal{V}_m\}$ de $\mathbb{R}^d$ en m régions de Voronoï :

$$\mathcal{V}_k = \left\{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x} - \boldsymbol{\mu}_k\| \leq \|\mathbf{x} - \boldsymbol{\mu}_j\| \text{ pour tout } j \neq k\right\}.$$

Dans la suite, une *boule* désigne le sous-ensemble de X contenu dans une région de

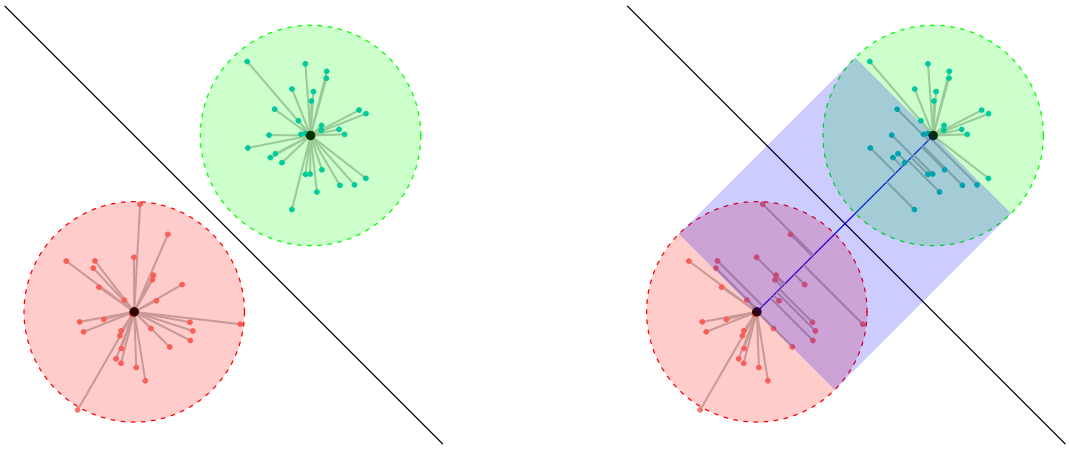


Figure 1: Boules (gauche) versus pont (droite). L'inertie de chaque structure est la somme des distances au carré représentées par les lignes grises.

Voronoï. L'inertie de deux boules $\mathcal{V}_k$ et $\mathcal{V}_l$ est définie par :

$$I_{kl} = \sum_{\boldsymbol{x}_i \in \mathcal{V}_k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2 + \sum_{\boldsymbol{x}_i \in \mathcal{V}_l} \|\boldsymbol{x}_i - \boldsymbol{\mu}_l\|^2.$$

Nous définissons un *pont* comme une structure correspondant au segment reliant deux centroïdes $\boldsymbol{\mu}_k$ et $\boldsymbol{\mu}_l$. L'inertie d'un pont entre $\mathcal{V}_k$ et $\mathcal{V}_l$ est donnée par :

$$B_{kl} = \sum_{\boldsymbol{x}_i \in \mathcal{V}_k \cup \mathcal{V}_l} \|\boldsymbol{x}_i - \boldsymbol{p}_{kl}(\boldsymbol{x}_i)\|^2,$$

où

$$\boldsymbol{p}_{kl}(\boldsymbol{x}_i) = \boldsymbol{\mu}_k + t_i(\boldsymbol{\mu}_l - \boldsymbol{\mu}_k),$$

avec

$$t_i = \min\left(1, \max\left(0, \frac{\langle \boldsymbol{x}_i - \boldsymbol{\mu}_k | \boldsymbol{\mu}_l - \boldsymbol{\mu}_k \rangle}{\|\boldsymbol{\mu}_l - \boldsymbol{\mu}_k\|^2}\right)\right).$$

En considérant deux centroïdes, la moyenne normalisée de la différence entre l'inertie du pont et celle des boules constitue la base de notre mesure d'affinité entre deux régions :

$$\begin{aligned} \frac{B_{kl} - I_{kl}}{(n_k + n_l)\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|^2} &= \frac{\sum_{\boldsymbol{x_i} \in \mathcal{V}_k} \langle \boldsymbol{x_i} - \boldsymbol{\mu}_k | \boldsymbol{\mu}_l - \boldsymbol{\mu}_k \rangle_+^2 + \sum_{\boldsymbol{x_i} \in \mathcal{V}_l} \langle \boldsymbol{x_i} - \boldsymbol{\mu}_l | \boldsymbol{\mu}_k - \boldsymbol{\mu}_l \rangle_+^2}{(n_k + n_l)\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|^4}, \\ &= \frac{\sum_{\boldsymbol{x_i} \in \mathcal{V}_k \cup \mathcal{V}_l} \alpha_i^2}{n_k + n_l}, \end{aligned}$$

où

$$\alpha_i = \begin{cases} t_i, & \text{si } t_i \in [0, 1/2], \\ 1 - t_i, & \text{si } t_i \in]1/2, 1]. \end{cases}$$

L'idée principale derrière cette affinité est que t_i représente la position relative de la projection de $\boldsymbol{x}_i$ sur le segment $[\boldsymbol{\mu}_k, \boldsymbol{\mu}_l]$. La valeur α_i représente la position relative sur le segment, avec comme point de référence le centroïde de la classe à laquelle $\boldsymbol{x}_i$ appartient.

Cette quantité peut également être interprétée en relation avec une forme locale de l'ACP. Contrairement au critère d'inertie standard de l'ACP, cette approche est directionnelle et limitée à l'union de deux cellules de Voronoï.

Soit $X_{|k}$ la matrice de données restreinte à la région $\mathcal{V}_k$. L'opérateur de projection sur le segment centré en $\boldsymbol{\mu}_k$ et dirigé vers $\boldsymbol{\mu}_l - \boldsymbol{\mu}_k$ est donné par :

$$\boldsymbol{\pi}_{kl}(\cdot) = \boldsymbol{p}_{kl}(\cdot) - \boldsymbol{\mu}_k.$$

On en déduit alors :

$$\frac{\sum_{\boldsymbol{x_i} \in \mathcal{V}_k \cup \mathcal{V}_l} \alpha_i^2}{n_k + n_l} = \frac{\|\boldsymbol{\pi}_{kl}(X_{|k})\|_F^2 + \|\boldsymbol{\pi}_{lk}(X_{|l})\|_F^2}{(n_k + n_l)\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|^2},$$

où $\|\cdot\|_F$ est la norme de Frobenius.

Cette formulation est particulièrement utile pour l'implémentation numérique.

De plus, l'hyperplan $\mathcal{P}_{kl}$ qui sépare les deux clusters définis par les centroïdes $\boldsymbol{\mu}_k$ et $\boldsymbol{\mu}_l$ est perpendiculaire au segment reliant les centroïdes et coupe ce segment en son milieu.

Si l'on considère tous les points $\boldsymbol{x}_i \in \mathcal{V}_k \cup \mathcal{V}_l$ qui ne sont pas projetés sur les centroïdes mais quelque part sur le segment, alors la distance d'un point à l'hyperplan est donnée par :

$$d(\boldsymbol{x}_i, \mathcal{P}_{kl}) = (1/2 - \alpha_i)\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|.$$

Cette distance est similaire au concept de marge en Machines à Vecteurs de Support (*Support Vector Machines*, SVM) (Cortes and Vapnik 1995). Lorsque les valeurs de α_i sont faibles (proches de zéro, car $\alpha_i \in [0, 1/2]$), les marges par rapport à l'hyperplan sont grandes, ce qui indique une faible densité entre les classes. À l'inverse, si les marges sont petites, cela suggère que les deux classes appartiennent à une même région densément peuplée. Par conséquent, la somme des α_i ou α_i^2 augmente avec la densité de la région entre les classes. Notons que ce critère est local et indique la différence relative de densité entre les boules et le pont, plutôt que d'évaluer un score global des densités des structures.

Finalement, nous définissons l'affinité du pont entre les centroïdes k et l comme la racine carrée du gain de variance :

$$a_{kl} = \begin{cases} 0, & \text{si } k = l, \\ \sqrt{\frac{\sum_{\boldsymbol{x_i} \in \mathcal{V}_k \cup \mathcal{V}_l} \alpha_i^2}{n_k + n_l}}, & \text{sinon.} \end{cases}$$

L'inclusion de la racine carrée redéfinit la mesure d'affinité basée sur la variance. Plutôt que d'utiliser la norme euclidienne au carré, l'affinité est interprétée comme une moyenne quadratique, représentant le rapport entre l'écart-type et la longueur du segment reliant deux centroïdes.

Afin de permettre aux points avec une grande marge de dominer et de rendre l'algorithme plus robuste au bruit et aux valeurs aberrantes, nous considérons la transformation exponentielle $\tilde{a}_{kl} = g(a_{kl}) = \exp(\gamma a_{kl})$, où γ est un facteur d'échelle. Ce facteur est défini de manière à garantir une séparation suffisante entre les coefficients finaux. Il est déterminé par l'équation :

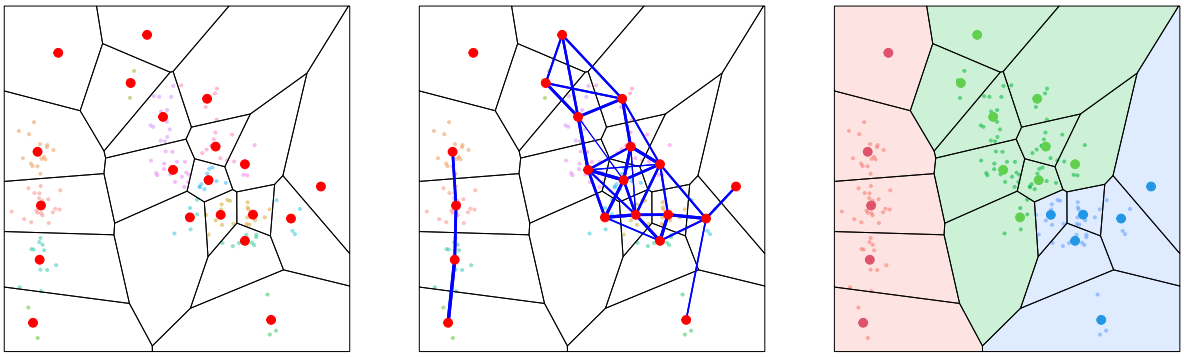
$$\gamma = \frac{\log(M)}{q_{90} - q_{10}},$$

où q_{10} et q_{90} sont respectivement les 10e et 90e percentiles de la matrice d'affinité originale, et $M > 1$. Ainsi, étant donné que la transformation préserve l'ordre, le 90e percentile de la nouvelle matrice est M fois plus grand que le 10e percentile. Par défaut, M est arbitrairement fixé à une valeur élevée de 10^4. Cette transformation réduit les valeurs maximales de la matrice d'affinité tout en préservant l'interprétabilité de la métrique et ses propriétés proches d'une distance.

3 Algorithme

Algorithm 1 Spectral Bridges

procedure SpectralBridges(X, k, m)
  centroids, voronoiRegions ← KMeans(X, m)
  $A = \{g(a_{kl})\}_{kl}$ ← Affinity(X, centroids, voronoiRegions)
  labels ← SpectralClustering(A, k)
  clusters ← Propagate(X, labels, voronoiRegions)
  return clusters
end procedure



(a) Vector quantization (b) Affinity computation (c) Spectral clustering

Figure 2: Illustration de l'algorithme Spectral Bridges avec le jeu de données Iris (premier plan principal). Les points rouges en gras représentent les centroïdes des clusters, tandis que les cellules colorées indiquent la partition finale des données. (a) Quantification vectorielle, (b) Calcul d'affinité, (c) Clustering spectral et propagation.

Bibliographie

Arthur, David and Sergei Vassilvitskii (June 2006). *k-means++: The Advantages of Careful Seeding*. Technical Report 2006-13. Stanford InfoLab. url: http://ilpubs.stanford.edu:8090/778/.

Carlsson, Gunnar (2009). "Topology and Data". In: *Bulletin of the American Mathematical Society* 46, pp. 255–308. doi: 10.1090/S0273-0979-09-01249-X.

Chazal, Frédéric, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba (2013). "Persistence-based clustering in Riemannian manifolds". In: *Journal of the ACM (JACM)* 60.6, pp. 1–38.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

Dhillon, Inderjit S, Yuqiang Guan, and Brian Kulis (2004). "Kernel k-means, spectral clustering and normalized cuts". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 551–556.

Eisen, Michael B., Paul T. Spellman, Patrick O. Brown, and David Botstein (1998). "Cluster analysis and display of genome-wide expression patterns". In: *Proceedings of the National Academy of Sciences* 95.25, pp. 14863–14868.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *kdd*. Vol. 96, pp. 226–231.

Laplante, Félix and Christophe Ambroise (2024). "Spectral Bridges". In: *Computo*.

Latouche, Pierre, Etienne Birmelé, and Christophe Ambroise (2011). "Overlapping stochastic block models with application to the French political blogosphere". In: *The Annals of Applied Statistics* 5.1, pp. 309–336. doi: 10.1214/10-AOAS382. url: https://doi.org/10.1214/10-AOAS382.

MacQueen, James et al. (1967). "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. Oakland, CA, USA, pp. 281–297.

McLachlan, Geoffrey J. and David Peel (2000). *Finite Mixture Models*. New York: Wiley-Interscience. isbn: 9780471006268.

Mur, Angel, Raquel Dormido, Natividad Duro, Sebastian Dormido-Canto, and Jesús Vega (2016). "Determination of the optimal number of clusters using a spectral clustering optimization". In: *Expert Systems with Applications* 65, pp. 304–314. issn: 0957-4174. doi: https://doi.org/10.1016/j.eswa.2016.08.059. url: https://www.sciencedirect.com/science/article/pii/S0957417416304572.

Shi, Jianbo and Jitendra Malik (2000). "Normalized cuts and image segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8, pp. 888–905.

Von Luxburg, Ulrike (2007). "A tutorial on spectral clustering". In: *Statistics and computing* 17, pp. 395–416.

Classification non supervisée sous contraintes strictes tranchées

Christian Derquenne

Chercheur indépendant - chris.emr@wanadoo.fr

Résumé. Cet article se place dans le cadre de la classification non supervisée pour détecter des structures complexes dans les données (formes allongées, formes emboitées, densités différentes, ...). Pour cela, nous proposons un critère sous contraintes strictes tranchées. Tout d'abord, chacune des variables est normalisée entre 0 et 1 sur lesquelles la distance de Manhattan est appliquée individuellement, puis une matrice de dissimilarité binaire est construite à l'aide d'un seuil de voisinage. Enfin, le critère du plus proche voisin est appliqué sur celle-ci afin d'obtenir un nombre de classes à seuil fixé. La détection du ou des nombre(s) de groupes potentiels est réalisé à l'aide du dendrogramme binaire associé.

Mots-clés. Classification non supervisée, structures complexes, dissimilarités binaires, densité

Abstract. This article is placed in the framework of unsupervised clustering to detect complex structures in data (elongated shapes, nested shapes, different densities, ...). For this, we propose a criterion under strict constraints. First, each of the variables is normalized between 0 and 1 on which the Manhattan distance is applied individually, then a binary dissimilarity matrix is constructed using a neighborhood threshold. Finally, the nearest neighbor criterion is applied on it in order to obtain a number of clusters with a fixed threshold. The detection of the number(s) of potential groups is carried out using the associated binary dendrogram.

Keywords. Unsupervised clustering, complex structures, stepwise methods, binary dissimilarities, density

1 Contexte - objectif

Préambule : Cette modeste contribution est en hommage à Pierre Cazes.

Dans de nombreuses applications, la recherche d'une structure dans les données à l'aide de la détection de groupes (patterns) représente un des objectifs primordiaux (recherche des formes d'agrégats d'étoiles pour modéliser l'évolution des galaxies, détection d'actions frauduleuses en finance, détection de comportements anormaux dans la foule lors d'événements sportifs, regroupement de courbes de consommations d'électricité pour établir des tarifs adaptés, ...). Les deux grands types d'approches proposés dans la littérature sont probabilistes et exploratoires. Le premier suppose une connaissance de la loi a priori des données, alors que le second type n'y est pas contraint. Ce papier se focalise sur les approches exploratoires de classification non supervisée. Chacune d'elles repose sur des critères d'optimisation pour la recherche du nombre "adéquat" de classes. Cette tâche est d'autant plus ardue que la forme et le contenu de celles-ci ne sont pas connus non plus à l'avance. Dans les cas très simples, elles aboutissent généralement toutes aux mêmes résultats (bon nombre de classes et bons contenus associés)

comme le montre la ligne 5 de la figure 1 [Scikit-Learn, 2023]. Par contre, les structures complexes comme des formes emboitées (lignes 1 et 2), des formes allongées (ligne 4), des formes de dispersions hétérogènes (ligne 3), ... que l'oeil détecte sans problème, seules quelques méthodes les découvrent (Spectral Clustering, DBSCAN et OPTICS, dans ces exemples). L'objectif de l'approche proposée nommée "classification non supervisée sous contraintes strictes tranchées" est de découvrir des patterns présentant des structures complexes. Dans ce papier, nous décrirons et discuterons brièvement des avantages et des inconvénients des méthodes concurrentes. Puis, nous détaillerons l'approche proposée sur un exemple et nous comparerons ses résultats aux méthodes. Nous conclurons sur les apports et les faiblesses de l'approche proposée, et nous fournirons quelques voies futures en termes d'amélioration et de nouveaux développements.

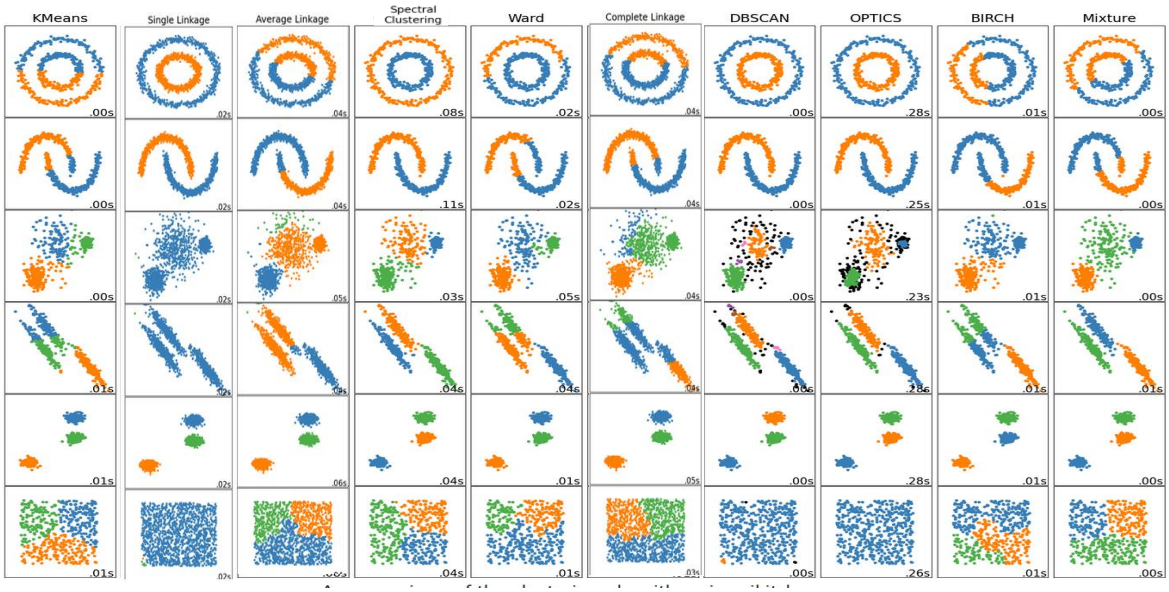


Figure 1 : Résultats des critères de classification non supervisée

2 Proposition d'une approche de classification non supervisée sous contraintes strictes tranchées

Comme indiqué dans l'introduction, la découverte de la "vraie" structure dans les données est très complexe, voire encore impossible même avec les méthodes les plus puissantes développées actuellement. En effet, de nombreux challenges sont à relever sur l'ensemble des étapes de la classification non supervisée : choix de la transformation ou non des données d'entrée, choix de la ou des distances appropriées entre les données, choix du ou des critères de classification qui paraissent adéquats sachant les forces et les faiblesses de ceux-ci sur telles ou telles structures, choix de ou des indices de détection du "bon" nombre de classes et de qualité de la typologie (compacité et isolation des classes, globale), adaptabilité à la grande dimension (nombre d'individus et nombre de variables), interprétation des classes.

2.1 Discussion sur les critères de classification non supervisée

Les méthodes de classification non supervisée exploratoires suivent principalement quatre stratégies : hiérarchique (Ward, single et complete linkage, average linkage), à nombre de classes fixé (k-means, par exemple), densité (DBSCAN, OPTICS) et factorielle (Spectral Clustering).

Les critères hiérachiques utilisent la matrice de dissimilarités entre les objets (individus, classes : le nombre d'individus doit donc être raisonnable, mais le nombre de variables peut être être élevé car il entre dans le calcul des dissimilarités) et procèdent par agrégations successives sans remise en cause de la partition précédente. Le choix du ou des nombres de classes potentiel(s) est effectué à partir d'une différence élevée entre deux valeurs successives du critère d'agrégation. Le critère du plus proche voisin est généralement adapté à des classes de formes non triviales (concentriques, emboitées, allongées, ...), à condition qu'elles soient suffisamment éloignées. Ce critère peut entraîner des effets de chaînage (cf. fig. 1). L'outil visuel du dendrogramme est généralement utilisé pour aider au choix du nombre de classes.

La méthode des k-means repose sur la minimisation de l'inertie intra-classe et permet de traiter de grosses bases données, mais l'utilisateur doit choisir un nombre de classes a priori. L'algorithme débute par le choix d'un nombre de noyaux égal à celui du nombre de classes fixé, puis rassemble autour de chaque noyau ses individus les plus proches au sens de la distance euclidienne. A l'itération suivante, les nouveaux noyaux correspondent aux plus proches du centre de gravité des nouvelles classes, l'algorithme se poursuit jusqu'à convergence. Les k-means ne permettent pas de découvrir des classes de formes complexes (fig. 1).

L'idée principale des approches fondées sur la densité est d'identifier les régions dans l'espace des variables où les points de données sont denses, puis d'extraire des classes en fonction de ces régions, sans a priori sur le nombre de classes. Ce type d'approches permet généralement de découvrir des classes de formes complexes. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [Esther et al., 1996], algorithme couramment utilisé, regroupe les points de données proches les uns des autres en fonction d'une mesure de distance et d'un seuil de densité. Pour cela, il est nécessaire de fixer un voisinage (rayon et/ou nombre de voisins) autour de chaque objet. La taille de celui-ci est la clé de la réussite pour obtenir une typologie "optimale" ce qui engendre de nombreux réglages. De plus, DBSCAN a des difficultés à détecter des classes de densités différentes. OPTICS (Ordering Points To Identify Clustering Structure) [Ankers et al., 1999] est une extension de DBSCAN qui détecte des classes de densités différentes et utilise la classification hiérarchique des rayons de voisinage. Mais OPTICS présente deux inconvénients, le coût mémoire et l'accroissement du temps de calcul avec la complexité. DBSAN et OPTICS découvrent l'ensemble des formes (fig. 1) et affichent la présence de points noirs hors-classes.

La classification spectrale [Cristianini et al., 2001] est un algorithme de partitionnement des données reposant sur la théorie spectrale des graphes et sur l'algèbre linéaire. L'idée est de segmenter un graphe en plusieurs petits groupes ayant des valeurs similaires ou proches. L'algorithme se déroule en plusieurs étapes : la construction d'un graphe à l'aide d'une matrice de similarités, la segmentation des points de données dans des espaces dimensionnels plus petits (matrice laplacienne) et, l'utilisation des valeurs propres et vecteurs propres pour définir les sous-graphes (le nombre de valeurs propres nulles correspond au nombre de classes). Puis la méthode des k-means est appliquée sur les vecteurs propres de cette matrice. La classification spectrale extrait bien les classes (fig. 1).

2.2 Le critère de classification sous contraintes strictes et tranchées

La genèse de cette approche vient d'une étude visant à découvrir, parmi plusieurs compétitions de para-cyclisme sur route, l'existence d'une structure de densité de course. Pour atteindre cet objectif, nous avons construit et classifié des indicateurs de densité de course sur l'ensemble des compétitions. Ces indicateurs étaient distribués sur une seule dimension, mais leur classification était compliquée même visuellement. Nous avons développé notre propre méthode.

Le cas unidimensionnel. La première étape de cette approche consiste à normaliser les données entre 0 et 1, telle que : $\tilde{X} = (X - \min(X))/(\max(X) - \min(X))$. La deuxième calcule un seuil θ représentant une distance entre les individus. Un choix naturel est l'équirépartition des n objets, tel que $\theta = 1/(n-1)$. La troisième étape revient à construire de proche en proche les classes. Comme les individus sont ordonnés de façon croissante sur une seule dimension, nous commençons par calculer la distance $d(1,2)$ entre $x_{(1)} = \min(X)$ et $x_{(2)}$. Si $d(1,2) < \theta$ alors $x_{(1)}$ et $x_{(2)}$ sont rassemblés dans la même classe C_1, sinon $x_{(1)}$ constitue à lui seul la classe C_1 et $x_{(2)}$ est le premier élément de C_2. Le processus continue : une nouvelle classe C_m se forme si $d(i-1,i) \geq \theta$, tel que $x_{(i-1)} \in C_{m-1}$. L'algorithme se termine lorsque les n individus ont été parcourus. A la fin, nous aurons M classes constituées de n_m objets, tel que $\sum_{m=1}^{M} n_m = n$. Nous pouvons constater que le choix de la valeur du seuil θ est strict, il pourrait être adouci en augmentant sa largeur à l'aide d'un paramètre α, tel que $\theta(\alpha) = 1/(n-1)(1+\alpha)$ où α varie entre 0 et $n-2$, ($\theta \subset [0;1]$). Cette procédure "à la main" permet de construire une matrice de dissimilarités binaires $\mathbf{D}_b$, telle que chaque élément est défini par $d_b(i,i') = 1_{[d(i,i') \geq \theta(\alpha)]}$. Une matrice adjacente peut être construite, telle que $\mathbf{S}_b = 1 - \mathbf{D}_b$. La quatrième étape consiste à fournir $\mathbf{D}_b$ en entrée du critère du plus proche voisin. Le choix du nombre de classes est déduit du dendrogramme qui devient binaire. En effet, les hauteurs des branches sont toutes identiques et le nombre de branches correspond au nombre de classes. Par ailleurs, la matrice adjacente $\mathbf{S}_b$ permet de construire un graphe et pourra être utilisée en entrée de la classification spectrale.

Le cas multidimensionnel. Comme dans le cas unidimensionnel, chaque variable X_j, $j = 1, ..., p$ est normalisée sur $[0;1]$. La deuxième étape consiste à calculer une matrice $\mathbf{D}_j$ de distances (par exemple, Manhattan) pour chaque variable, une valeur $\theta(\alpha)$ est fixée. Ce qui permet de déduire la matrice binaire de similarités $\mathbf{S}_{b(j)}$. En rassemblant ces p matrices, nous obtenons une nouvelle matrice comme suit : $\mathbf{S}_b = \sum_{j=1}^{p} \mathbf{S}_{b(j)}$. La diagonale de $\mathbf{S}_b$ est composée d'éléments qui valent tous p, alors que les éléments triangulaires varient entre 0 et p. Plus un élément $s_b(i,i')$ se rapproche de p, plus les objets i et i' sont proches sur l'ensemble des dimensions. Cette matrice peut être normalisée, telle que $\tilde{\mathbf{S}}_b = \mathbf{S}_b/p \subset [0;1]$, ce qui est le propre d'une matrice de similarités. La contrainte stricte dans le nom de ce critère signifie que seuls les éléments égaux à 1 sont retenues. Cela permet de construire une nouvelle matrice de similarités binaires strictes : $\mathbf{S}_{bs}$, telle que chaque élément $s_{bs}(i,i') = 1_{[\tilde{s}_b(i,i')=1]}$. Une autre possibilité pour obtenir $s_{bs}(i,i')$ est d'utiliser le max des distances, tel que : $s_{bs}(i,i') = 1_{[\max_j d_j(i,i') < \theta(\alpha))]}$. Ces deux matrices adjacentes sont identiques. Enfin, à partir de celles-ci nous tirons la matrice de dissimilarités binaires strictes : $\mathbf{D}_{bs} = 1 - \mathbf{S}_{bs}$. Par conséquent, $\mathbf{D}_{bs}$ pourra être utilisée en entrée du critère du plus proche voisin, alors que $\mathbf{S}_{bs}$ sera fournie en entrée de la classification spectrale. Enfin, le terme "tranchée" dans le nom du critère proposé revient à relacher la contrainte stricte en introduisant la notion de tranche. Pour cela, au lieu d'imposer à $s_b(i,i')$ d'être égale à p pour

déduire $s_{bs}(i,i') = 1$, il est possible de réduire la contrainte à $s_b(i,i') = p-1$. Une tranche représente simplement un degré de liberté retiré en passant de p à $p-1$. En d'autres termes, une tranche correspond à une dimension dans l'espace des données distribuées sur p variables. Cependant cette contrainte relachée peut se révéler trop stricte. Dans ces conditions, nous proposons d'introduire des sous-tranches $slice_t(\alpha)$ de taille $1/(n-2-\alpha)$ avec $t = 1, n-2-\alpha$. Lorsque $t = n-2-\alpha$, cela est équivalent à imposer $s_b(i,i') = p-1$, sinon $s_b(i,i') = p - slice_t(\alpha)$.

Comment choisir le nombre de classe ?. Le paramètre α joue un rôle crucial dans la recherche du nombre de classes et de la qualité de la typologie. Nous avons développé des indicateurs permettant de juger la qualité de la typologie pour chaque valeur de α. Pour chacune de celles-ci, un nombre de classes est proposé. L'idée la plus naturelle est de mesurer le gain ou la perte obtenus entre deux typologies en α et $\alpha+1$. Parmi les nombreux indicateurs que nous avons testés, deux ont retenu notre attention. Ils sont les plus pessimistes possible, afin de ne garder que les meilleures typologies complexes car ils sont fondés sur la plus petite distance entre classes. Le premier indice $d_{min}(\alpha)$ (1) tient compte de la plus petite distance entre deux individus appartenant à deux classes distintes sur les p variables. L'évolution de cet indice est monotone croissante sur $[0;1]$ en fonction de α. Un saut important lors du passage de α à $\alpha+1$ peut révéler un nombre potentiel de classes, tel que : $M^*(d_{min}(\alpha+1)) = \arg\max_{\alpha}(d_{min}(\alpha+1) - d_{min}(\alpha))$. Le même nombre de classes peut être proposé pour plusieurs α successifs. Le second indicateur (2) intègre le seuil $\theta(\alpha)$) de façon à mesurer son écart avec le premier. Il varie entre 0 et 1. Le calcul d'un nombre potentiel de classes $M^*(\gamma)$ suit le même raisonnement que celui de d_{min}.

$$d_{min}(\alpha) = \min_{j=1,pm=1,M-1} \min_{k=2,M} \min_{\substack{i\in C_m \\ i'\in C_k}} d_j(i,i') \qquad (1) \qquad\qquad \gamma(\alpha) = (d_{min}(\alpha) - \theta(\alpha))/(p - \theta(\alpha)) \qquad (2)$$

2.3 Application du critère et comparaison

Nous détaillons le déroulement des étapes du critère proposé sur des données réparties en trois cercles concentriques (fig. 5). Le plus petit cercle contient 50 objets ; le cercle moyen possède 75 individus et le plus grand regroupe 125 éléments. Tout d'abord, la matrice de dissimilarités (fig. 2) de type Manhattan est visualisée sous forme d'une cartographie de couleurs dégradées de jaune à marron ; l'intensité est d'autant plus claire que la dissimilarité est faible. Nous pouvons constater la présence de 9 gros rectangles. Les trois carrés sur la diagonale possédent une ligne blanche épaisse correpondant aux objets liés dans la même classe. Les figures 3a et 3b représentent les évolutions des deux indicateurs sur lesquelles quatre classes semblent le choix optimal (le plus grand saut) pour $\alpha = 13$ ($\theta = 0,056$) et, $d_{min}(\alpha = 13) = 0,108$ et $\gamma(\alpha = 13) = 0,027$. En sélectionnant cette valeur fixée, nous pouvons afficher le dendrogramme (fig. 4) issu du critère du plus proche voisin. Le nombre de classes est déduit du nombre de hautes branches identiques. La figure 5 affiche le résultat de la méthode. Quatre classes sont bien identifiées, le cercle bleu est proche d'un point extérieur (numéro 183). La méthode n'a pas englouti cet objet et a pu détecter parfaitement les classes. La table 1 fournit les indicateurs par classe. La classification spectrale a été appliquée sur la matrice de dissimilarités de Manhattan. Seule la dernière valeur propre est égale à 0, donc une seule classe, ce qui signifie que ce critère n'a pas pu détecter la structure typologique. Le critère du plus proche voisin aboutit au même résultat que la méthode proposée, ainsi que DBSAN et OPTICS.

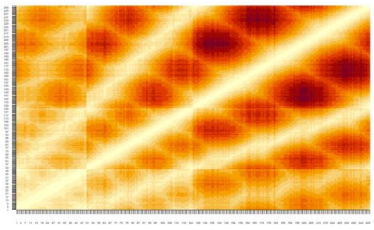


Figure 2 : Cartographie des dissimilarités

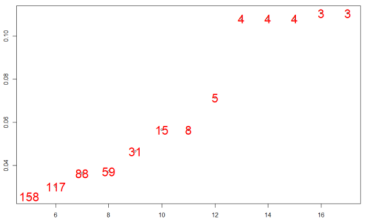


Figure 3a : Evolution de $d_{min}(\alpha)$

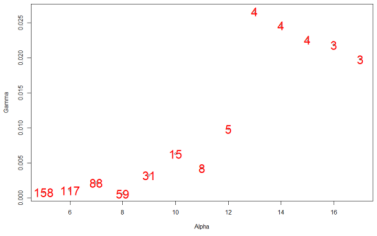


Figure 3b : Evolution de $\gamma(\alpha)$

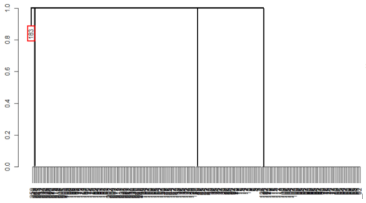


Figure 4 : Dendrogramme Single Linkage

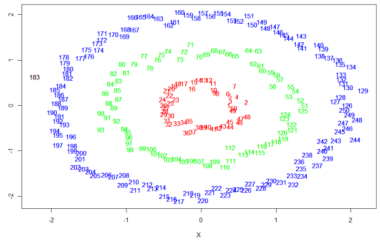


Figure 5 : Résultats du critère

m	$d_{min}^{(m)}(\alpha)$	$\gamma^{(m)}(\alpha)$	n_m
1	0,112	0,059	50
2	0,110	0,057	75
3	0,108	0,055	124
4	0,108	0,055	1

Table 1 : Indicateurs par classe

3 Apports, applications et voies futures

Le nouveau critère de classification proposé permet de détecter des classes de structures différentes (formes concentriques, emboitées, allongées, ...). Seuls DBSCAN et OPTICS fournissent des résultats comparables sur de nombreux jeux de données testés. Un des autres avantages de cette approche est de proposer une liste restreinte de classes potentiellement optimales par rapport au critère du plus proche voisin dont le choix est souvent difficile. L'utilisation de la contrainte stricte permet d'obtenir des groupes compactes quelle que soit leur forme et de détecter des valeurs atypiques comme DBSCAN et OPTICS. L'effet de chainage du critère du plus proche voisin devient un avantage lorsqu'il est utilisé sur la matrice de dissimilarités binaires. Le point faible de la méthode proposée est probablement la taille des données (nombre d'individus, notamment). Ce critère a fourni de très bons résultats sur des données réelles en cosmologie, en sport, en marketing, ... La réduction de la complexité de l'algorithme, le développement de nouveaux indicateurs de qualité de la typologie, la généralisation de cette méthode à des variables de nature différente et l'extension à la classification de variables constituent les voies futures.

Bibliographie

Ankerst M., Breunig M., Kriegel H. P. and & Sander J., (1999), OPTICS: ordering points to identify the clustering structure, *ACM Sigmod record 28*, **2**, 49-60.
Cristianini N., Shawe-Taylor J. & Kandola J., (2001), Spectral kernel methods for clustering. In Proceedings of Neural Information Processing Systems, volume 14.
Ester M., Kriegel H. P., Sander J. & Xu X., (1996), A density-based algorithm for discovering clusters in large spatial databases with noise, *KDD*, **96**(34), 226-231.

Segmentation d'images spectrales de fluorescence X sur la base d'une dissimilarité s'appuyant sur la perte de vraisemblance

Maxime Megel[1] & Gilles Celeux [2] & Serge Cohen [1] & Agnès Grimaud [3] & Christine Keribin [4]

[1] *Université Paris-Saclay, CNRS, ministère de la culture, UVSQ, MNHN, IPANEMA, 91192 Gif-sur-Yvette, France, maxime.megel@cnrs.fr et serge.cohen@cnrs.fr*

[2] *Université Paris-Saclay, Inria, Laboratoire de mathématiques d'Orsay, France gilles.celeux@inria.fr*

[3] *Université Paris-Saclay, UVSQ, CNRS, Laboratoire de mathémathiques de Versailles, 78000, Versailles, France, agnes.grimaud@uvsq.fr*

[4] *Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, France christine.keribin@universite-paris-saclay.fr*

Résumé. L'imagerie par fluorescence de rayon-X (XRF) génère des images spectrales de grande taille, enregistrant un spectre d'énergie en chaque pixel. Elle est utilisée en particulier dans l'étude des matériaux anciens afin de déterminer systématiquement des zones homogènes, la fluorescence de rayon-X permettant de déterminer la composition élémentaire des matériaux. Le *clustering* de telles données est délicat à cause de leur volumétrie, aussi bien en nombre de pixels qu'en nombre d'énergie d'émission, mettant à mal des méthodes comme la classification hiérarchique. Dans un précédent travail, nous avons proposé une classification hiérarchique sous contrainte de proximité utilisant une dissimilarité basée sur la distance euclidienne pondérée du χ^2 entre les spectres. Mais cette méthode présentait des artefacts manifestes. Dans la présente communication, nous poussons l'exploitation du modèle posant l'émission de photons XRF par une loi de Poisson en chaque canal. Ceci nous permet de proposer une dissimilarité basée sur une différence de log-vraisemblance et un critère de pseudo-inertie lié à une déviance. Cette nouvelle approche a été appliquée sur des jeux de données réelles et simulées. Les résultats obtenus montrent que ce critère est bien adapté à la segmentation hiérarchique d'images XRF et surpasse en qualité la méthode utilisant la distance du χ^2.

Mots-clés. Segmentation, image spectrale, clustering hiérarchique, dissimilarité, déviance, fluorescence de rayon X.

Abstract. X-ray fluorescence (XRF) imaging generates large spectral images, recording an energy spectrum at each pixel. It is used in the study of ancient materials to systematically determine homogeneous areas whose XRF determines the elemental composition of the materials. Clustering of such data is tricky because of their volume due to the number of pixels as well as the number of emission energies, undermining methods such as hierarchical clustering. In a previous work, we proposed a hierarchical classification under proximity constraint using a dissimilarity based on a weighted Euclidean distance of χ^2 between spectra. But this method presented apparent artifacts. In this communication, we take a modeling

approach by representing the XRF photon emission by Poisson distributions. This model allows us to propose a dissimilarity based on a difference in log-likelihood and a pseudo-inertia criterion linked to deviance. This new approach is applied to real and simulated datasets. The results obtained show that this criterion is well adapted to the hierarchical clustering of XRF images and outperforms the method using the χ^2 distance.

Keywords. Segmentation, spectral imaging, hierarchical clustering, dissimilarity, deviance, X-ray fluorescence

1 Introduction

La fluorescence de rayon X (XRF) est une méthode largement utilisée pour déterminer la composition élémentaire de matériaux, en mesurant leur spectre d'émission de photons. Elle permet de dresser une cartographie XRF de l'objet d'étude en balayant le flux de rayon X, et produit une image dont chaque pixel est associé au spectre d'émission de la zone balayée. Cette technique est souvent utilisée dans l'étude des matériaux anciens, car ils sont mal connus a priori et hétérogènes. Actuellement, les instruments de mesure permettent de collecter des images de XRF constituées de plusieurs millions de pixels et à chacun des pixels est associé un spectre de quelques milliers de canaux d'énergie. Ainsi la richesse des informations collectées est stockée dans des fichiers de plusieurs dizaines de GO, trop volumineux pour utiliser des algorithmes standards de segmentation. Il est donc nécessaire de développer des algorithmes d'analyse performants en coût de calcul et permettant l'exploitation des connaissances a priori sur les données.

[Celeux et al., 2022] ont proposé une méthode de segmentation par classification hiérarchique ascendante intégrant des contraintes spatiales : la distance euclidienne du χ^2 est choisie pour dissimilarité, l'agrégation se fait suivant le critère de Ward et intègre une contrainte garantissant de ne fusionner lors des premières étapes que les clusters spatialement adjacents, ce qui préserve la structure de l'image en ne proposant que des classes simplement connexes. Mais cette méthode souffre de quelques limitations, par exemple une robustesse limitée lors de l'augmentation du niveau de bruit des mesures, c'est-à-dire avec un temps plus court d'exposition du matériau aux rayons X. La qualité des résultats peut également souffrir d'une trop grande hétérogénéité entre les différentes zones d'un même échantillon, ou encore de problèmes d'ombrage quand la morphologie de l'échantillon crée de très fortes variations de l'intensité totale d'un pixel à l'autre. Ainsi, nous proposons une nouvelle méthode dont l'un des buts est de mieux gérer les variations d'intensité, que ce soit d'un pixel à l'autre ou entre différents canaux de chacun des spectres, grâce à une dissimilarité plus robuste face à des niveaux élevés de bruit.

Dans la section 2, revenant sur le principe physique d'émission, nous proposons une modélisation probabiliste des spectres à l'aide de distributions de Poisson pour décrire le comportement aléatoire des photons détectés. Cela nous permet, dans la section 3, de construire un nouveau critère de dissimilarité, basé sur la perte de vraisemblance lors de l'agrégation de région. Il est alors possible d'utiliser des critères de sélection de modèle tel que l'ICL

afin de déterminer un nombre pertinent de clusters. Enfin dans la section 4, cette nouvelle approche est testée sur des jeux de données simulées et réelles et permet d'améliorer les résultats obtenus précédemment.

2 Modèle

Une image spectrale est constituée de n pixels et à chaque pixel $i \in \{1, ..., n\}$ est associé un spectre de dimension M, la m-ième coordonnée correspondant au nombre de photons détectés dans le canal d'énergie m ($1 \leq m \leq M$).

2.1 Spectre d'un pixel

En chaque pixel i, le processus d'émission et de captation des photons est aléatoire. Soit $p_i(m)$ la probabilité qu'un photon incident provoque une détection dans le canal de fréquence m. Le vecteur $\mathbf{p_i} = (p_i(m))_{m=1}^M$ est le *profil spectral* en i et vérifie $\sum_{m=1}^M p_i(m) = 1$.

Lors de l'acquisition, le faisceau incident sur la zone i amène la production stochastique de photons mesurés avec une intensité β_i. Le nombre de photons détectés N_i suit une loi de Poisson de paramètre β_i : $N_i \sim \text{Poisson}(\beta_i)$, les photons se répartissant sur les différentes énergies m ($1 \leq m \leq M$) en fonction du profil spectral. Ainsi, nous observons en chaque pixel i le vecteur aléatoire $\mathbf{S_i} = (S_i(m))_{m=1}^M$, appelé *spectre d'émission*, de loi de probabilité $\mathbf{S_i} \sim \text{Spec}(\beta_i, \mathbf{p_i})$, telle que:

$$\mathbb{P}\left[\mathbf{S_i} = \mathbf{s_i}\right] = e^{-\beta_i} \beta_i^{n_i} \prod_{m=1}^{M} \frac{(p_i(m))^{s_i(m)}}{s_i(m)!}.$$

C'est la loi d'un vecteur dont les marginales sont indépendantes, de loi $S_i(m) \sim \text{Poisson}(\beta_i p_i(m))$.

2.2 Spectre d'un ensemble de pixels

Pour l'ensemble de l'image qui contient n pixels, on dispose de n spectres qui sont indépendants du fait de la nature stochastique du phénomène observé. À chaque sous-ensemble de pixels $C \subset \{1, ..., n\}$, on attribue le *spectre somme* correspondant $\mathbf{S_C} := \sum_{i \in C} \mathbf{S_i}$.
Soit N_C, le nombre de photons détectés correspondant:

$$N_C \sim \text{Poisson}\left(\sum_{i \in C} \beta_i\right) \text{ et } S_C(m) \sim \text{Poisson}\left(\sum_{i \in C} \beta_i p_i(m)\right).$$

En notant $\beta_C = \sum_{i \in C} \beta_i$ et $p_C(m) = \sum_{i \in C} \beta_i p_i(m) / \beta_C$, on déduit: $\mathbf{S_C} \sim \text{Spec}(\beta_C, \mathbf{p_C})$. Notons que les caractéristiques d'un spectre somme de deux classes disjointes C_1 et C_2 se déduisent facilement lors de leur agrégation :

$$\mathbf{S_C} = \mathbf{S_{C_1}} + \mathbf{S_{C_2}};\ \ N_C = N_{C_1} + N_{C_2};\ \ \beta_C = \beta_{C_1} + \beta_{C_2};\ \ p_C(m) = \frac{\beta_{C_1} p_{C_1}(m) + \beta_{C_2} p_{C_1}(m)}{\beta_{C_1} + \beta_{C_2}}$$

3 Agrégation hiérarchique

L'objectif est de segmenter l'image en régions de composition homogène sans connaître à l'avance le nombre de classes pertinentes pour l'échantillon étudié. La tâche est donc essentiellement non supervisée. Le but est de développer un algorithme performant en temps de calcul, prenant en compte la taille volumineuse des jeux de données manipulés et permettant de préserver les matériaux anciens en limitant leur temps d'exposition aux rayons X. Pour cela, nous partons de l'algorithme proposé par [Celeux et al., 2022] combinant dans un premier temps la technique de classification hiérarchique ascendante avec des contraintes spatiales avant de relâcher cette contrainte spatiale lors des dernières itérations. L'originalité de la nouvelle méthode proposée est le remplacement du critère d'agrégation initial, la distance du χ^2, par la définition d'une nouvelle dissimilarité basée sur une différence de vraisemblance posant la comparaison entre spectres comme une question d'homogénéité entre des échantillons d'ensemble de photons caractérisés par leur énergie.

3.1 La perte de vraisemblance comme dissimilarité

La log-vraisemblance $\mathcal{L}(\ .\ |\mathbf{S_i})$ associée à un spectre $\mathbf{S_i}$ s'écrit :

$$\mathcal{L}(\mathbf{p_i}, \beta_i|\mathbf{S_i}) = \sum_{m=1}^{M} S_i(m)\log(p_i(m)) + N_i\log(\beta_i) - \beta_i - \sum_{m=1}^{M}\log(S_i(m)!)$$

Par indépendance des spectres, la log-vraisemblance de deux spectres est la somme de leurs log-vraisemblances :

$$\mathcal{L}(\mathbf{p_i}, \beta_i, \mathbf{p_j}, \beta_j|\mathbf{S_i}, \mathbf{S_j}) = \mathcal{L}(\mathbf{p_i}, \beta_i|\mathbf{S_i}) + \mathcal{L}(\mathbf{p_j}, \beta_j|\mathbf{S_j})$$

et on note sa valeur maximale $\widehat{\mathcal{L}} := \mathcal{L}(\widehat{\mathbf{p}_\mathbf{i}}, \widehat{\beta}_i, \widehat{\mathbf{p}_\mathbf{j}}, \widehat{\beta}_j|\mathbf{S_i}, \mathbf{S_j})$. Restreignant maintenant l'optimisation à la contrainte d'égalité des profils spectraux $\mathbf{p_i} = \mathbf{p_j} = \mathbf{p_{ij}}$, on définit la vraisemblance maximale sous cette contrainte : $\widehat{\mathcal{L}_0} := \mathcal{L}(\widehat{\mathbf{p_{ij}}}, \widehat{\beta}_i, \widehat{\mathbf{p_{ij}}}, \widehat{\beta}_j|\mathbf{S_i}, \mathbf{S_j})$. Nous définissons alors la dissimilarité entre deux spectres comme la perte de log-vraisemblance entre l'hypothèse de deux profils spectraux différents contre celle d'un même profil :

$$d^2_{\mathcal{L}}(\mathbf{S_i}, \mathbf{S_j}) := \widehat{\mathcal{L}} - \widehat{\mathcal{L}_0}.$$

puis on étend cette définition à la dissimilarité entre classes.

Définition 3.1. *Soient deux classes C_k et C_ℓ d'intersection vide et leurs spectres somme respectifs $\mathbf{S_{C_k}}$ et $\mathbf{S_{C_\ell}}$. La dissimilarité $d_{\mathcal{L}}(,)$ par perte de vraisemblance est définie par :*

$$\begin{aligned} d^2_{\mathcal{L}}(\mathbf{S_{C_k}}, \mathbf{S_{C_\ell}}) = &\sum_{m=1}^{M} S_{C_k}(m)\log(S_{C_k}(m)) + S_{C_\ell}(m)\log(S_{C_\ell}(m)) \\ &- N_{C_k}\log(N_{C_k}) - N_{C_\ell}\log(N_{C_\ell}) + (N_{C_k} + N_{C_\ell})\log(N_{C_k} + N_{C_\ell}) \\ &- \sum_{m=1}^{M}(S_{C_k}(m) + S_{C_\ell}(m))\log(S_{C_k}(m) + S_{C_\ell}(m)) \end{aligned}$$

avec la convention $0\log(0) = 0$.

Cette définition s'applique à la dissimilarité entre les spectres $\mathbf{S_i}$ et $\mathbf{S_j}$ en considérant les deux classes $C_k = \{i\}$ et $C_\ell = \{j\}$.

3.2 Représentant de classe et pseudo-inertie

Traditionnellement, le représentant est le centre de gravité. Cette notion se traduit mal dans le cas de la vraisemblance. En revanche, les propriétés du spectre somme (*cf.* section 2.2) assurent que ce dernier est un représentant naturel de la classe.

L'inertie d'une classe est définie comme la somme (éventuellement pondérée) des carrés des distances au centre de gravité de la classe. Cette définition ne peut être utilisée ici, le représentant de la classe n'étant pas le centre de gravité. Nous pouvons toutefois construire une notion de pseudo-inertie qui permet de retrouver les propriétés qui fondent la méthode de Ward : positivité, séparation (la pseudo-inertie est nulle si tous spectres sont égaux), conservation de la pseudo-inertie totale et cohérence de la variation d'inertie.

Définition 3.2. *La pseudo-inertie intra-classe d'un sous-ensemble C est définie par :*

$$I_C = \left[\sum_{i\in C}\left[\sum_{m=1}^{M} S_i(m)\log(S_i(m))\right] - N_i\log(N_i)\right] - \left[\sum_{m=1}^{M} S_C(m)\log(S_C(m))\right] + N_C\log(N_C)$$

C'est la différence de log-vraisemblance entre le modèle saturé définissant un profil spectral par pixel de C contre le modèle à un profil spectral commun à tous les pixels de C. On peut le voir comme une déviance nulle.

Reprenant la même idée, on définit la pseudo-inertie interclasse comme la différence de log-vraisemblance entre un profil spectral pour chacun des K représentants contre un même profil pour ces K représentants. On note $\mathbf{S_{tot}}$ le spectre associé à la classe $C = \{1, ..., n\}$ contenant tous les pixels et N_{tot} le nombre de photons détectés dans cette classe.

Définition 3.3. *La pseudo-inertie interclasse d'une segmentation $\mathcal{C}^K = (C_k)_{k=1}^K$ est*

$$I_e(\mathcal{C}^K) = \left[\sum_{k=1}^{K}\left[\sum_{m=1}^{M} S_{C_k}(m)\log(S_{C_k}(m))\right] - N_{C_k}\log(N_{C_k})\right] - \left[\sum_{m=1}^{M} S_{tot}(m)\log(S_{tot}(m))\right] + N_{tot}\log(N_{tot})$$

On vérifie la conservation de la pseudo-inertie, analogue du théorème de Huyghens : pour toute segmentation $\mathcal{C}^K = (C_k)_{k=1}^K$: $I_a(\mathcal{C}^K) + I_e(\mathcal{C}^K) = I_{tot}$, où I_{tot} est la pseudo-inertie intra classe quand tous les pixels sont dans la même classe $C = \{1, ..., n\}$. De plus, la variation de pseudo-inertie interclasse lors d'une agrégation est égal à la diminution du carré de la dissimilarité entre les deux classes fusionnées.

Ces propriétés permettent ainsi de reprendre l'algorithme proposé par [Celeux et al., 2022] en y remplaçant la dissimilarité du χ^2 par celle par perte de vraisemblance $d_{\mathcal{L}}(,)$.

3.3 Sélection du nombre de clusters

Le fait d'utiliser une modélisation probabiliste permet d'utiliser des critères de sélection de modèle pour le choix du nombre de clusters, plutôt que de s'en remettre à des critères empiriques comme la règle du coude ou le saut maximal d'inertie. Les critères BIC [Schwarz, 1978] et ICL [Biernacki et al., 2000] sont fréquemment utilisés. Si le BIC est un critère d'identification, consistant dans le cas de modèle de mélange [Keribin, 2000], ICL met aussi l'accent sur la qualité du clustering généré, ce qui est notre objectif. Son expression s'écrit ici :

$$ICL(\mathcal{M}^K) = KM\log(n) - 2\sum_{k=1}^{K}\left[|C_k|\log|C_k| - N_{C_k}\log N_{C_k} + \sum_{m=1}^{M} S_{C_k}(m)\log S_{C_k}(m)\right]$$

4 Etude sur des jeux de données

A partir de jeux de données d'images spectrales de fluorescence X, nous montrerons, pour la segmentation par classification hiérarchique, l'apport de l'utilisation de la dissimilarité s'appuyant sur la perte de vraisemblance par rapport à la dissimilarité du χ^2 précédemment utilisée. En particulier, sur un jeu de données simulées, l'erreur de classification des pixels est moindre avec la dissimilarité par perte de vraisemblance, et ceci à divers niveaux de bruit. Dans le cas d'images spectrales réelles, l'utilisation de la dissimilarité par perte de vraisemblance permet de réduire les erreurs manifestes constatées avec la dissimilarité χ^2, faisant de la classification hiérarchique avec cette nouvelle dissimilarité un outil de choix pour la segmentation spectrale par fluorescence X.

References

[Biernacki et al., 2000] Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.

[Celeux et al., 2022] Celeux, G., Cohen, S. X., Grimaud, A., and Gueriau, P. (2022). Hierarchical clustering of spectral images with spatial constraints for the rapid processing of large and heterogeneous data sets. *SN Computer Science*, 3(3):194.

[Keribin, 2000] Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 49–66.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Forêts aléatoires non supervisées pour le clustering

Alice Simon [1] & Robin Genuer [1] & Boris Hejblum [1]

[1] *Univ. Bordeaux, INSERM, INRIA, BPH, U1219, F-33000 Bordeaux, France*
alice.simon@u-bordeaux.fr, robin.genuer@u-bordeaux.fr, boris.hejblum@u-bordeaux.fr

Résumé. Le clustering, ou classification non-supervisée, désigne un ensemble de méthodes d'analyse visant à identifier des structures ou des groupes naturels au sein d'un jeu de données. En regroupant les données suivant leurs similitudes intrinsèques, il constitue un outil clé pour l'exploration et la segmentation de vastes bases de données. Dans ce contexte, nous proposons `RFCLUST`, une nouvelle méthode de clustering inspirée du principe général de la méthode des forêts aléatoires (Breiman, 2001) en apprentissage supervisé. La méthode `RFCLUST` est donc obtenue en agrégeant un ensemble d'arbres, reposant chacun sur un clustering hiérarchique divisif monothétique (Chavent et al., 2007), qui sépare les données en deux sous-groupes à chaque nœud de l'arbre en optimisant un critère d'inertie. L'ensemble des arbres ainsi construits permet d'obtenir une matrice de co-clustering, représentant la probabilité que chaque paire d'observations appartienne au même cluster. Cette matrice permet de quantifier l'incertitude entourant la partition des observations. D'autre part, elle peut ensuite être exploitée par des méthodes de partitionnement classiques afin d'obtenir les clusters finaux, par exemple en appliquant une Classification Ascendante Hiérarchique par la méthode de Ward. `RFCLUST` conserve les mécanismes clés des forêts aléatoires, notamment le rééchantillonnage par bootstrap des observations et la sélection aléatoire de variables. Cela permet une robustesse et une diversité dans la construction des partitions tout en préservant l'interprétabilité des arbres individuels. Nous avons étudié l'impact des paramètres du modèle, notamment mtry et le nombre de feuilles de chaque arbre, sur le jeu de données `palmerpenguins`. Ce dernier contient des mesures morphologiques de trois espèces de manchots sur trois îles de l'archipel Palmer, en Antarctique.

Mots-clés. Apprentissage non supervisé, Forêts aléatoires, Clustering Hiérarchique

Abstract. Clustering, or unsupervised classification, refers to a set of analytical methods aimed at identifying natural structures or groups within a dataset. By grouping data based on their intrinsic similarities, it serves as a key tool for exploring and segmenting large-scale databases. In this context, we introduce `RFCLUST`, a novel clustering method inspired by the principle of random forests (Breiman, 2001) used in supervised learning. `RFCLUST` is constructed by aggregating an ensemble of trees, where each tree is built using a monothetic divisive hierarchical clustering approach (Chavent et al., 2007). At each node, data are split into two subgroups by optimizing an inertia criterion. The resulting set of trees produces a co-clustering matrix, which encodes the probability that each pair of observations belongs to the same cluster. This matrix not only provides a way to quantify the uncertainty associated with the clustering process, but can also be leveraged by classical partitioning methods, such as Ward's method (Agglomerative Hierarchical Clustering), to obtain the final clusters.

RFCLUST retains key mechanisms from random forests, such as bootstrap resampling of observations and random selection of variables, ensuring robustness and diversity in partitioning while preserving the interpretability of individual trees. We investigated the impact of model parameters, particularly mtry and the number of leaves per tree, using the palmerpenguins dataset. This dataset comprises morphological measurements of three penguin species collected from three islands in the Palmer Archipelago, Antarctica.

Keywords. Unsupervised Learning, Random Forests, Hierarchical Clustering

1 Introduction

Les forêts aléatoires (Breiman, 2001) constituent une méthode performante d'apprentissage supervisé. Cette approche repose sur les arbres de décision de type CART (Classification and Regression Trees), tout en corrigeant leurs limites grâce à deux stratégies : d'une part, un échantillonnage aléatoire avec remplacement des observations pour chaque arbre ; d'autre part, une sélection aléatoire sans remplacement des variables à chaque nœud de l'arbre.
Jusqu'à présent, les forêts aléatoires ont été peu étendues pour l'apprentissage non supervisé. Bien que certaines méthodes proposent déjà de générer des données synthétiques (Shi and Horvath, 2006; Afanador et al., 2016) et exploitent les forêts aléatoires pour distinguer ces données des observations réelles, elles ne sont pas spécifiquement conçues pour le clustering. Or, la construction même des forêts aléatoires leur confère une robustesse et une capacité d'exploration qui peuvent être des atouts considérables dans ce cadre. C'est dans cette optique que des méthodes comme `sidClustering` (Mantero and Ishwaran, 2021) ont été développées. Celle-ci exploite le plein potentiel des forêts aléatoires pour le clustering, au prix toutefois de transformations complexes des variables en amont.
De notre côté, nous proposons `RFCLUST`, une nouvelle méthode qui substitue aux traditionnels arbres CART l'algorithme `DIVCLUS-T` (Chavent et al., 2007), un algorithme de clustering hiérarchique divisif monothétique.

2 Méthodes

2.1 `DIVCLUS-T` : Clustering Hiérarchique Divisif Monothétique

L'algorithme `DIVCLUS-T`, pour `DIV`isive `CLUS`tering `T`ree, scinde progressivement les données en sous-groupes homogènes tout en garantissant une interprétation explicite des groupements obtenus. Contrairement aux méthodes de clustering hiérarchique agglomératif qui fusionnent progressivement les observations, `DIVCLUS-T` adopte une approche divisive : il démarre avec l'ensemble des observations regroupées dans un seul et même cluster puis procède à des divisions successives.
Sa particularité réside dans son caractère monothétique : à chaque étape, la séparation est effectuée sur la base d'une seule variable explicative. Cette propriété permet de lire le dendrogramme obtenu en sortie comme un arbre de décision, offrant ainsi une vision claire des

critères de partitionnement.
Enfin, bien que la méthode soit divisive, elle repose sur le même principe d'optimisation que certaines approches classiques du clustering : chaque séparation est réalisée en minimisant l'inertie intra-cluser, à l'image du clustering hiérarchique agglomératif de Ward ou du partitionnement par k-means.
L'algorithme 1 présente les principales étapes de la méthode DIVCLUS-T.

Algorithm 1 DIVCLUS-T : Clustering Hiérarchique Divisif Monothétique

Entrée : Ensemble de données $\mathcal{D} = \{x_1, \ldots, x_n\}$, nombre de clusters souhaité K
Sortie : Partition $\mathcal{P}$ de $\mathcal{D}$ en K clusters
Initialisation : $\mathcal{P} \leftarrow \{\mathcal{D}\}$
while $\text{card}(\mathcal{P}) < K$ **do**
    for chaque cluster $C \in \mathcal{P}$ **do**
        Initialiser $h(C) \leftarrow 0$
        for chaque variable X^j dans $\{X^1, \ldots, X^p\}$ **do**
            for chaque question binaire candidate Q sur X^j **do**
                Définir la partition $(A, \bar{A})$ de C induite par Q
                Calculer l'inertie inter-cluster associée B_Q :

$$B_Q = \frac{\text{card}(A)\,\text{card}(\bar{A})}{\text{card}(A) + \text{card}(\bar{A})}\, d_M^2\Big(g(A),\, g(\bar{A})\Big)$$

▷ où $g(U)$ correspond au centroïde du cluster U

                if B_Q est meilleur que le meilleur score actuel pour C **then**
                    Mettre à jour : $Q_C \leftarrow Q$, $h(C) \leftarrow B_Q$
                    Enregistrer la partition optimale : $(A_C, \bar{A}_C) \leftarrow (A, \bar{A})$
                end if
            end for
        end for
    end for
    Sélectionner le cluster $C^* \in \mathcal{P}$ tel que

$$C^* = \arg\max_{C \in \mathcal{P}} h(C)$$

    Diviser C^* en deux sous-clusters en utilisant sa meilleure bipartition $(A_{C^*}, \bar{A}_{C^*})$
    Mettre à jour la partition :

$$\mathcal{P} \leftarrow \Big(\mathcal{P} \setminus \{C^*\}\Big) \cup \{A_{C^*}, \bar{A}_{C^*}\}$$

end while
return $\mathcal{P}$

2.2 RFCLUST : Random Forests à partir de DIVCLUS-T

L'algorithme DIVCLUS-T fournit donc un cadre efficace pour structurer les données de manière hiérarchique et monothéthique, propice à la création d'une nouvelle méthode de clustering inspirée du principe des forêts aléatoires en apprentissage supervisé.
Nous proposons ainsi RFCLUST, qui repose sur l'agrégation d'un ensemble d'arbres, où chaque arbre est construit en utilisant l'algorithme DIVCLUS-T. À l'instar des forêts aléatoires en apprentissage supervisé, RFCLUST introduit de la diversité en appliquant un rééchantillonnage par bootstrap des observations et une sélection aléatoire de variables ($mtry$) à chaque arbre de la forêt. Cette combinaison permet d'atténuer la variabilité des partitions et de capturer différentes structures présentes dans les données. L'ensemble des arbres ainsi générés sert ensuite à construire une matrice de co-clustering, qui estime la probabilité que chaque paire d'observations appartienne à un même cluster à travers les différentes partitions individuelles associées aux arbres. Cette matrice peut alors être exploitée par la suite par des méthodes de partitionnement classiques afin d'obtenir les clusters finaux.
L'algorithme 2 décrit le fonctionnement de notre méthode RFCLUST lorsque le nombre de clusters K est fixé a priori. Toutefois, RFCLUST peut également s'affranchir de cette contrainte en estimant automatiquement le nombre optimal de clusters pour chaque arbre, à partir du critère de Ward.

Algorithm 2 RFCLUST : Random Forests à partir de DIVCLUS-T

Entrée : Ensemble de données $\mathcal{D} = \{x_1, \ldots, x_n\}$, nombre d'arbres souhaité $ntrees$, nombre de clusters souhaité K des arbres individuels, nombre de variables tirées aléatoirement $mtry$
Sortie : Matrice de similarité $\mathcal{M}$
for chaque arbre de la forêt **do**
    Réaliser un bootstrap $\mathcal{B}$ de $\mathcal{D}$ avec remise
    Réaliser un tirage aléatoire $var_{\mathcal{B}}$ de $mtry$ variables sans remise
    Appliquer DIVCLUS-T aux éléments de $\mathcal{B}$ expliqués par les variables $var_{\mathcal{B}}$ en fixant le nombre de clusters souhaité à K
    Remplir & Stocker la matrice de similarité de ce bootstrap $\mathcal{M}_{\mathcal{B}}$
end for
Calculer $\mathcal{M}$ en moyennant les matrices de similarités en fonction du nombre d'apparitions de chaque paire d'observations sur les $ntrees$ bootstraps
return $\mathcal{M}$

3 Résultats

Nous avons évalué les performances de notre nouvelle méthode de clustering sur le jeu de données palmerpenguins (Horst et al., 2020). Cette base de données contient 167 individus, appartenant à trois espèces distinctes de manchots mâles de l'archipel Palmer, en Antarctique. Ces trois espèces constituent les trois clusters que l'algorithme est censé retrouver. Chaque individu est décrit par quatre variables numériques caractérisant ses traits morphologiques.

L'algorithme `RFCLUST` a été appliqué sans spécifier a priori le nombre de groupes à identifier. Pour chaque arbre de la forêt, le nombre optimal de clusters a été déterminé de manière automatique à partir du critère de Ward. L'analyse a été menée sur une forêt de 500 arbres, où les quatre variables numériques ont été mises en compétition à chaque noeud de chaque arbre.
La matrice de similarité obtenue en sortie de la forêt est représentée sous forme de heatmap en figure 1.

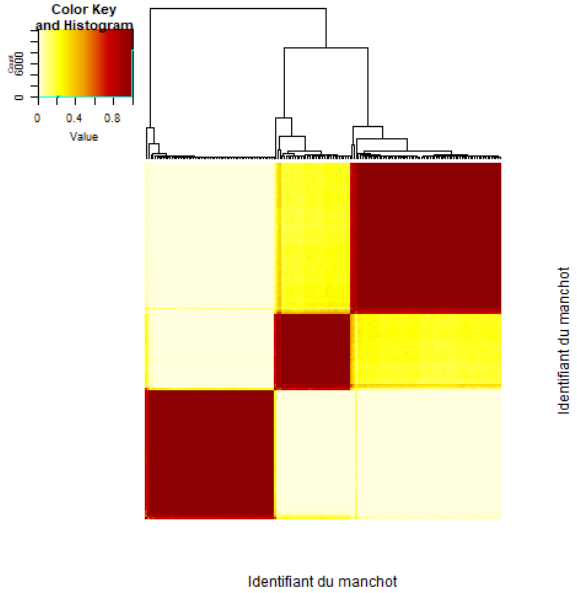


FIGURE 1 – Matrice de similarité obtenue sur les données `palmerpenguins` en sortie de `RFCLUST` pour une forêt de 500 arbres. Le dendrogramme a lui été obtenu en appliquant la méthode de Ward (Clustering Ascendant Hiérarchique) avec comme point de départ cette matrice de similarité.

La matrice de similarité met en évidence trois groupes bien distincts, comme le confirme le dendrogramme obtenu, suggérant que le choix d'une partition en trois classes est particulièrement pertinent. Cette observation est confirmée quantitativement par un Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) de 0.97, calculé à partir de la partition résultant de l'application de la méthode de Ward sur la matrice de similarité. Ces résultats confirment la capacité de `RFCLUST` à identifier efficacement la structure des données, sans connaissance préalable du nombre de classes.
Lorsque le paramètre *mtry* varie entre 1 et 4 dans ce contexte, la configuration $mtry = 1$ ne permet pas d'identifier clairement la structure en trois groupes dans la matrice de similarité obtenue. Effectivement, dans ce cas-là, chaque arbre est construit en se basant sur une unique variable numérique et une des trois espèces n'est pas séparée des autres.

4 Discussion

Dans ce travail, nous avons construit des forêts aléatoires en étendant la méthode de clustering hiérarchique divisif monothétique `DIVCLUS-T` de Chavent et al. (2007). La robustesse et la finesse d'exploration offertes grâce aux forêts aléatoires et l'interprétabilité des résultats de `DIVCLUS-T` semblent expliquer les performances encourageantes de `RFCLUST`. Nous avons également implémenté dans `RFCLUST` une calibration automatique du nombre de clusters optimal pour chaque arbre de la forêt en exploitant le critère de Ward. Cela a l'avantage de construire des partitions cohérentes suivant un critère quantitatif contrairement aux méthodes nécessitant un choix a priori du nombre de clusters finaux.

Cependant, pour le moment, pour chaque arbre de notre forêt, nous choisissons un sous-ensemble de *mtry* variables avant la construction de l'arbre et non à chaque nœud de celui-ci comme cela est le cas dans les forêts aléatoires classiques. Nous prévoyons donc de faire évoluer la construction de l'arbre `DIVCLUS-T` dans ce sens en espérant gagner en capacité d'exploration.

Pour le moment, notre méthode s'appuie sur une évaluation de la similarité entre les paires d'observations qui ne tient compte que de la partition finale de chaque arbre. Nous prévoyons d'utiliser plus finement la structure des arbres individuels pour définir une autre notion de distance entre nos paires d'observations.

Références

Afanador, N. L., Smolinska, A., Tran, T. N., and Blanchet, L. (2016). Unsupervised random forest : a tutorial with case studies. *journal of Chemometrics*, 30(5) :232–241.

Breiman, L. (2001). Random forests. *Machine learning*, 45 :5–32.

Chavent, M., Lechevallier, Y., and Briant, O. (2007). Divclus-t : A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis*, 52(2) :687–701.

Horst, A. M., Hill, A. P., and Gorman, K. B. (2020). *palmerpenguins : Palmer Archipelago (Antarctica) penguin data.* R package version 0.1.0.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2 :193–218.

Mantero, A. and Ishwaran, H. (2021). Unsupervised random forests. *Statistical Analysis and Data Mining : The ASA Data Science Journal*, 14(2) :144–167.

Shi, T. and Horvath, S. (2006). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1) :118–138.

Clustering sur graphes et clustering multi-vues

Fusing Multiple Data Views with Diffusion Maps

Gwendal Debaussart* & Argyris Kalogeratos

Centre Borelli, ENS Paris-Saclay, Université Paris-Saclay, 91190 Gif-sur-Yvette, France

Résumé. Les progrès dans l'acquisition de données ont permis la création de jeu de données multi-vues, où plusieurs point-de-vue décrivent de mêmes individus. Ces vues peuvent en pratique correspondre à différentes modalités, plusieurs instruments de mesures, ou encore à plusieurs transformations d'un jeu de donnée originel. Une des questions principales dans ce cadre est alors celle de la fusion d'information : comment peut-on déduire une unique représentation qui allie les informations comprises dans les différentes vues. Pour répondre à ce questionnement, nous introduisons une méthodologie inspirée des Diffusion Maps, en construisant un processus de diffusion à partir des différentes vues ; une telle méthodologie admet une interprétation probabiliste et géométrique assez naturelle. Nous arrivons à montrer qu'une telle méthodologie généralise certaines méthodes existantes dans ce cadre d'application. Nous démontrons son efficacité dans le cadre de partitionnement de données, à la fois pour des jeux de données synthétique mais également pour des données réelles.

Mots-clés. multi-vues, diffusion maps, apprentissage de représentation, apprentissage multi-modal, plongement, fusion de données, partitionnement.

Abstract. Progress in data acquisition has allowed the production of multi-view data, where each view provides a different representation of the same data objects. Those views may correspond to multiple data modalities (e.g text, images, sensor data ...), multiple measuring instrumentations, or may be produced via transformations over an original input data view. One of the main approach for multi-view representation is the fusion, where we aim at finding a common representation for the different views. To address this, we propose a methodology inspired by the Diffusion Maps framework, where we construct a diffusion process by making use of the different views. Such methodology admits natural probabilistic and geometric interpretation. We also demonstrate the fact that it encapsulate some of the current state of the art methods. We showcase the power of the proposed approach within the context of the data clustering task, through experiments on toy and real-world datasets.

Keywords. multi-view, diffusion maps, representation learning, multi-modal learning, manifold learning, embedding, data fusion, clustering.

*Corresponding author.
E-mail adress: gwendal.debaussart@ens-paris-saclay.fr

1 Introduction

Over the past decade, information fusion has emerged as one of the main challenges in the modern multi-view data representation field. In particular, it has been shown that instead of directly concatenating the different views into one, exploiting the multi-view aspect helps uncovering better representations of such datasets, notably in unsupervised problems like data clustering (e.g. see Papalexakis et al. (2013); Li et al. (2019); Chao et al. (2021)). Random-walk-based frameworks have been widely known and used in the single-view case. One example is the Diffusion Maps framework (Coifman and Lafon, 2006) that uses the random walk Laplacian to embed the data in a PCA-like fashion. The success of this approach has led to generalizations for the multi-view case, such as those by Lindenbaum et al. (2020); Kuchroo et al. (2022); Katz et al. (2019); Tang et al. (2020); Wang et al. (2012). As a fusion mechanism, Those methods use a general diffusion process that hinges on using the different diffusion processes associated to the different views.s

Contribution. We introduce a novel multi-view Diffusion Maps methodology that utilizes inhomogeneous random walks. Our framework encapsulates some of the state-of-the-art multi-view Diffusion Maps extensions. We show that, with some fine-tuning, this method yields promising results, comparing to the aforementioned existing alternatives.

2 Background

Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ be a single view input dataset, the Diffusion Maps framework (Coifman and Lafon, 2006) hinges on the creation of a random walk over the data-points. Let $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ be a symmetric kernel over the space of the datapoints, which allows for the computation of a similarity matrix $\mathbf{K} \in \mathbb{R}^{N\times N}$, where $\mathbf{K}_{ij} = \kappa(x_i, x_j)$. This matrix can then be normalized to become row-stochastic: by using its degree matrix $\mathbf{D}$, i.e. a diagonal matrix defined by $\mathbf{D}_{ii} = \sum_j \mathbf{K}_{ij}$, we can define the random walk transition matrix $\mathbf{P} = \mathbf{D}^{-1}K$. The main idea of the Diffusion Maps framework is to use the fact that this operator admits a real eigenvalue decomposition * that we can use to embed the datapoints. As $\mathbf{P}$ is a transition matrix, raising it to some power $t \in \mathbb{N}$ results in another transition matrix representing the random walk at a larger scale. As $\mathbf{P}^t$ admits a real eigenvalue decomposition, this defines a family of embeddings called *diffusion maps*. This has notably proven useful for denoising of datasets (Ángela Fernández et al., 2015; Kuchroo et al., 2022). In the context of multi-view learning, this procedure can be used for each view independently, to output a set of transition matrices $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2, \ldots\}$. Note that these operators are not necessarily created using the same similarity function, as the different views might have different geometries.

In the multi-view literature, most methods focus on either the fusion of similarity kernels (Lindenbaum et al., 2020) or the fusion of transition matrices (Katz et al., 2019; Kuchroo et al., 2022). In both cases, the main objective is to end up with a time-homogeneous Markov chain that can be iterated; we will denote its transition matrix by $\mathbf{P}'$. On the one hand, fusing

*this is due to the fact that $\mathbf{P} = \mathbf{D}^{-1}\mathbf{K}$, and $\mathbf{K}$ admits a real e.v decomposition as we assumed that it is a symmetric matrix.

the similarity kernels before normalization allows for the creation of operators that admits real eigenvalue decomposition (see Lindenbaum et al. (2020)). On the other hand, methods that focus on fusing the transition matrices allows for -generally- more computationally efficient and more direct fusion methodologies. As an example, Katz et al. (2019) compute $\mathbf{P}' = \mathbf{P}_1\mathbf{P}_2$, while Kuchroo et al. (2022) compute $\mathbf{P}' = \mathbf{P}_1^{t_1}\mathbf{P}_2^{t_2}$, where they give a way of tuning the time parameters $t_1, t_2 \in \mathbb{N}$. Unlike such existing works that focus on building pre-fixed composition of operators, we propose a methodology that offers more modeling flexibility.

3 Diffusion trajectories between views

Diffusion trajectories. Let $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2, \ldots\}$ be a set of transition matrices. We define a *diffusion trajectory* as a sequence $(\mathbf{W}_i)_{i\in\mathbb{N}} \in \mathcal{P}$ and the associated diffusion operator:

$$\mathbf{W}^{(t)} \triangleq \mathbf{W}_t\mathbf{W}^{(t-1)} \triangleq \mathbf{W}_t\mathbf{W}_{t-1}\ldots\mathbf{W}_1, \text{ where } \mathbf{W}^{(1)} \triangleq W_1.$$

The operator $\mathbf{W}^{(t)}$ encompasses some of the aforementioned methods. For the case of alternating diffusion (Katz et al., 2019), one can specify $\mathbf{W}^{(t)} = \mathbf{P}_1$ if t is odd, and $\mathbf{W}^{(t)} = \mathbf{P}_2$ if t is even. For the case of integrated diffusion (Kuchroo et al., 2022), one can specify $\mathbf{W}^{(t)} = \mathbf{P}_1$ if $t \leq t_1$, and $\mathbf{W}^{(t)} = \mathbf{P}_2$ if $t_1 < t \leq t_2$. To extend this process to $\mathbb{N}$, we can start by noticing that the identity matrix $\mathbf{I}_N$ is a stochastic matrix, and say that $\mathbf{W}^{(t)} = \mathbf{I}_N$ for any $t \geq t_2$.

Note that, defining $\mathcal{P}^t = \{\mathbf{W}^{(t)} \,|\, (\mathbf{W}_i)_{i\in\mathbb{N}} \in \mathcal{P}\}$ as the set of possible trajectories of length t, we can conceptually see $\cup_{t=1}^{\tau}\mathcal{P}^t$ as a tree, where the root node would be the identity matrix, and each node would correspond to a matrix belonging to $\mathcal{P}$. Then, we can interpret each node of this tree as a diffusion trajectory by looking its ancestors in the tree.

Trajectory-dependent Diffusion Maps. In general $\mathbf{W}^{(t)}$ is neither symmetric nor nearly being a symmetric matrix, and thus may admit complex eigenvalues. However, it can be seen that when $\mathcal{P}$ contains primitive matrices, $\mathbf{W}^{(t)}$ is also primitive, hence admits an ergodic distribution π, and the moduli of its eigenvalues are upper-bounded by 1. Multiple methods for dealing with transition matrices with complex eigenvalues can be considered, in this work, we choose to work with the singular value decomposition (SVD) as a work-around, let $\mathbf{W}^{(t)} = \mathbf{U}_t\mathbf{\Sigma}_t\mathbf{V}_t^\top$ be the SVD of $\mathbf{W}^{(t)}$. In this context, we can define the matrix:

$$\mathbf{\Psi}^{(t)} = \mathbf{U}_t\mathbf{\Sigma}_t \tag{1}$$

which is the diffusion map associated to the operator $\mathbf{W}^{(t)}$ of a trajectory. The embedding of the datapoint x_i is given by $\boldsymbol{e}_i^\top\mathbf{\Psi}^{(t)}$, where $\boldsymbol{e}_i$ is the i-th vector of the canonical basis of $\mathbb{R}^N$. In practice, one can use an algorithm such as the one given by Golub et al. (2000) for a quicker estimation of the SVD of the $\mathbf{W}^{(t)}$ matrix.

Selecting a trajectory. As stated previously, our multi-view diffusion trajectory framework opens to a wide variety of possibilities. For practical use, one has to search to find a 'good' trajectory for a problem at hand. Therefore, the introduction of this new space comprising all possible trajectories, allows for the definition of custom sampling pipelines that would seek for promising diffusion trajectories. In this work, we focus on the unsupervised learning

of those trajectories. As an example,in the data clustering context, we could associate an internal quality score to each possible trajectory, representing the quality of the clustering solution obtained by the diffusion maps related to the operator $\mathbf{W}^{(t)}$ of that trajectory. We obtain each such clustering using the k-means++ (Arthur and Vassilvitskii, 2007) algorithm on the datapoints represented by the truncated k-dimensional diffusion maps, where k is the number of clusters. Fig. 1 visualizes the clustering results for all possible trajectories of length $t = \{1, 2, 3\}$, as a tree graph with a circular layout. At each node, there appears a supervised evaluation score, namely the Adjusted Mutual Information (AMI) (higher score indicates better clustering result, as evaluated by ground truth labels). It can be observed

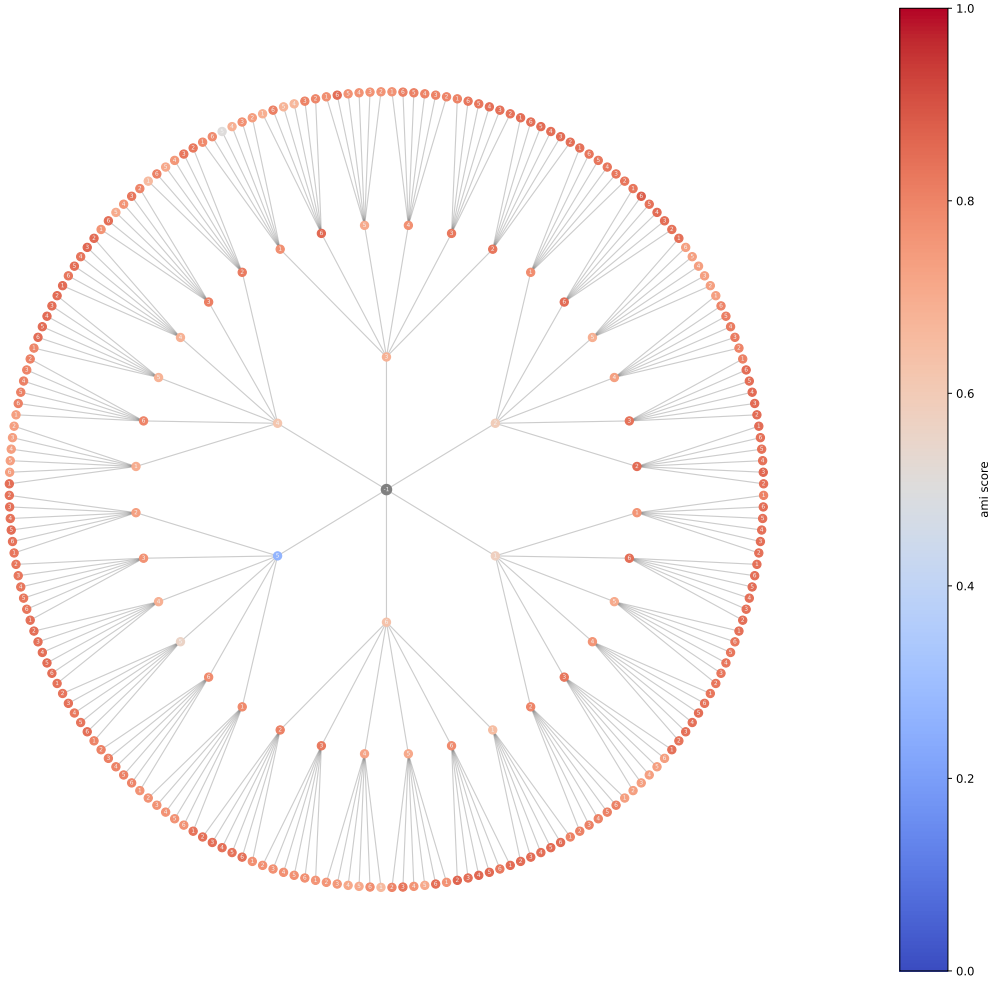


Figure 1: Visualization of the space $\cup_{t=1}^{3}\mathcal{P}^t$ associated to the Multiple-Feat dataset (6 views, 2000 datapoints) from the UCI repository. Each node corresponds to a specific trajectory-based diffusion operator $\mathbf{W}^{(t)}$, and its color represents the AMI score (more precisely, the mean AMI is over 100 experiments) of the clustering results obtained by the associated trajectory. Looking at the smooth way those colors are distributed over the tree graph, we can observe that some regions perform better than others.

that the graph seems smooth, in the sense that some regions of the space produce better results than others, while some others do not do so.

In practice, as implied earlier, it is not feasible to explore broadly such a search space. Instead, one could seek to optimize an unsupervised measure, for instance to maximize the Calinski-Harabasz index. There are various tools that could be employed for this optimization, e.g. tools of Global Optimization, such as the Direct (Jones, 2001) algorithm, or CMA-ES (Igel et al., 2007). As those tools generally rely on using continuous state spaces, $\mathscr{P}$ could be extended by constructing:

$$\overline{\mathscr{P}} = \left\{ \sum_{i=1}^{v} \alpha_i \mathbf{P}_i \,\middle|\, \mathbf{P}_i \in \mathscr{P}, \sum_i \alpha_i = 1 \right\}.$$

Note that, if all of elements of $\mathscr{P}$ are ergodic transition matrices, then all the elements of $\overline{\mathscr{P}}$ will also be transition matrices with that property. As the estimation of the singular values of the $\mathbf{W}^{(t)}$ is costly for large N, the search within the spaces $\mathscr{P}^t$ or $\overline{\mathscr{P}}^t$ should be frugal.

Utilization in a contrastive learning context. Making use of the intuition behind contrastive learning, one could aim at solving the following optimization problem, inspired by Pan and Kang (2021):

$$\min_{\substack{\mathbf{w}^{(t)} \in \overline{\mathscr{P}}, \\ \lambda_m : m=1,\dots,v}} \sum_{m=1}^{v} \lambda_m \left[\sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i^m} -\log \left(\frac{\exp(\mathbf{W}_{ij}^{(t)})}{\sum_{p \neq i}^{N} \exp(\mathbf{W}_{ip}^{(t)})} \right) \right],$$

where λ_m is a weight associated to the m-th view, such that $\sum_m \lambda_m = 1$, $\mathcal{N}_i^m$ represents the neighbors of the i-th node (i.e. corresponding to a datapoint) in the m-th view, and t is a fixed parameter. We suppose here that 'positive pairs' are nodes that belong in each other's nearest neighbors. Accordingly, note that the t parameter should not be taken too high, as it can be proven that $W^{(t)}$ converges to a rank one matrix almost surely (this result will be developed in a longer version of this work): when t is too big, the definition of 'positive' pairs based on $\mathbf{W}^{(t)}$ would then be misleading. For tuning t one could consider to rely on spectral quantities such as the singular entropy, i.e. the entropy of the singular values of $W^{(t)}$.

4 Conclusion

We have presented an extension of the Diffusion Maps framework for multi-view data. The main idea relies on the definition of a space of transition matrices, each element of which corresponding to a product of the transition matrices of the individual views. As this work is in early stage, we have included a visual example that shows the promise of our approach concerning the meaningfulness of searching within that space for finding a good diffusion trajectory for a problem at hand. Finally, he have identified some of the avenues that are possible for setting up a sampling pipeline and tuning of the procedure.

References

Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035.

Chao, G., Sun, S., and Bi, J. (2021). A Survey on Multiview Clustering. *IEEE Transactions on Artificial Intelligence*, 2(2):146–168.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30.

Golub, G., Solna, K., and Van Dooren, P. (2000). Computing the SVD of a general matrix product/quotient. *SIAM Journal on Matrix Analysis and Applications*.

Igel, C., Hansen, N., and Roth, S. (2007). Covariance matrix adaptation for multi-objective optimization. *Evolutionary Computation*.

Jones, D. R. (2001). *Direct global optimization algorithm*, pages 431–440. Springer US.

Katz, O., Talmon, R., Lo, Y.-L., and Wu, H.-T. (2019). Alternating diffusion maps for multimodal data fusion. *Information Fusion*, 45.

Kuchroo, M., Godavarthi, A., Tong, A., Wolf, G., and Krishnaswamy, S. (2022). Multimodal Data Visualization and Denoising with Integrated Diffusion.

Li, Y., Yang, M., and Zhang, Z. (2019). A Survey of Multi-View Representation Learning.

Lindenbaum, O., Yeredor, A., Salhov, M., and Averbuch, A. (2020). Multi-view diffusion maps. *Information Fusion*, 55:127–149.

Pan, E. and Kang, Z. (2021). Multi-view contrastive graph clustering. Issue: arXiv:2110.11842 arXiv: 2110.11842 [cs].

Papalexakis, E. E., Akoglu, L., and Ienco, D. (2013). Do more views of a graph help? Community detection and clustering in multi-graphs.

Tang, C., Liu, X., Zhu, X., Zhu, E., Luo, Z., Wang, L., and Gao, W. (2020). CGD: Multi-View Clustering via Cross-View Graph Diffusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5924–5931.

Wang, B., Jiang, J., Wang, W., Zhou, Z.-H., and Tu, Z. (2012). Unsupervised metric fusion by cross diffusion. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2997–3004.

Ángela Fernández, González, A. M., Díaz, J., and Dorronsoro, J. R. (2015). Diffusion maps for dimensionality reduction and visualization of meteorological data. *Neurocomputing*, 163:25–37.

Attributed Graph Partitioning: Comparing K-Means and Optimal Transport in Traffic Applications

Ioana Gavra [1] & Ketsia Guichard-Sustowski [2] & Loïc Le Marrec [2] & Véronique Thelen [3]

[1] *Univ Rennes 2, IRMAR - UMR 6625, ioana.gavra@univ-rennes2.fr*
[2] *Univ Rennes, IRMAR - UMR 6625, {ketsia.guichard, loic.lemarrec}@univ-rennes.fr*
[3] *Univ Rennes, CREM - UMR 6211, veronique.thelen@univ-rennes.fr*

Résumé. Le partitionnement d'un graphe attribué consiste à diviser les sommets en clusters distincts en tenant compte de la structure du graphe et des attributs des nœuds. L'objectif est d'adapter et de comparer deux méthodes de partitionnement de graphes attribués : k-means et une méthode issue du transport optimal. Cette dernière exploite une contrainte de masse relaxée pour déterminer des partitions sans prédéfinir le nombre de clusters. Nous étendons ces approches en combinant les distances topologiques et celles basées sur les attributs, afin de les appliquer aux graphes attribués. Un cadre de simulation permettra d'évaluer ces méthodes dans divers scénarios, en variant l'alignement entre les attributs des nœuds et les structures communautaires du graphe, ainsi que l'influence du bruit. Nous illustrerons ces méthodes avec des données réelles de mobilité. Les réseaux routiers peuvent en effet être modélisés comme des graphes avec des attributs (flux moyen de trafic, profils de vitesse), et partitionner ces réseaux facilite leur compréhension et gestion. Le choix de distances adaptées aux formes spécifiques de ces attributs (données fonctionnelles pour le trafic, histogrammes pour les vitesses) sera également discuté.

Mots-clés. graphes attribués, partitionnement de graphes, transport optimal, Gromov-Wasserstein, partitionnement de réseaux routiers

Abstract. The partitioning of an attributed graph involves dividing the vertices into distinct clusters, considering both the graph's structure and the node attributes. The goal is to adapt and compare two methods for partitioning attributed graphs: k-means and a method based on optimal transport. The latter exploits a relaxed mass constraint to determine the partition without predefining the number of clusters. We extend these approaches by combining topological distances with attribute-based distances to make them applicable to attributed graphs. A simulation framework will evaluate these methods in various scenarios, varying the alignment between node attributes and the graph's community structures, as well as the influence of noise. These methods will be illustrated with real mobility data. Road networks can indeed be modeled as graphs with attributes (e.g., average traffic flow, speed profiles), and partitioning these networks facilitates their understanding and management. The choice of distances suited to the specific forms of these attributes (functional data for traffic, histograms for speed) will also be discussed.

Keywords. attributed graph, graph partitioning, optimal transport, Gromov-Wasserstein, road network partitioning

1 Introduction

1.1 Context

Graph clustering aims to better understand large networks and discover communities within them. Furthermore, it might help identify vertices with a central position or specific edges between different clusters. This problem has broad applications across various fields, including social sciences, biology, and medicine (Fortunato 2010). Transportation is another domain where graph clustering can address a common challenge: road network partitioning. Managing complex and extensive road networks can be facilitated by dividing them into smaller, independent subnetworks, enabling the implementation of optimized control strategies.

Attributed graph clustering is at the confluence of graph clustering and multidimensional clustering of tabular data: we want to partition the graph into several clusters with a cohesive intra-cluster structure and homogeneous attribute values. In social network analysis, node attributes are typically discrete variables (such as age) or qualitative characteristics (e.g., gender or interests). However, in road network partitioning, attributes often take on more specific forms. Beyond length or qualitative characteristics, roads can be described by traffic metrics, such as average traffic flow (vehicles per time), average density (vehicles per length), or speed profiles. These traffic features are frequently represented as time series, functional data, or histograms.

Our objective is to adapt two methods to the context of attributed graphs and evaluate their performance using a dedicated simulation framework. We compare a distance-based method, k-means, adapted to this specific context, with an optimal transport-based partitioning approach. Simulated tests will enable a systematic comparison of methodologies across various criteria, such as scalability and sensitivity to noise.

1.2 Problem Description

We consider an undirected node-attributed graph as the triple $G = (V, E, A)$. $V = \{v_1, \ldots, v_n\}$ is the set of vertices, associated with a probability measure $\boldsymbol{\mu} \in \Sigma_n$, the probability simplex ($\Sigma_n := \{\boldsymbol{\mu} \in \mathbb{R}_+^n | \sum_i \mu(v_i) = 1\}$). $\boldsymbol{\mu} = \{\mu(v_1), \ldots, \mu(v_n)\}$ represents the probability distribution over the nodes, where $\mu(v_i)$ corresponds to the probability assigned to v_i. E is the set of edges, each edge $e_{ij} \in E$ is endowed with a length $l_{ij} > 0$. Finally, A is the set of attributes associated to the vertices. Attributes can be of various types, such as labels, continuous or discrete values. For the purposes of our application, we illustrate the method using functional attributes and histogram-based attributes.

2 Methods

Combe et al. (2012) propose storing both the structural and attribute information in a distance function between nodes, which can then be used with distance-based clustering methods:

$$d(v_i, v_j) = \alpha \cdot d_{\mathrm{S}}(v_i, v_j) + (1 - \alpha) \cdot d_{\mathrm{A}}(v_i, v_j)$$

where $d_{\mathrm{S}}(v_i, v_j)$ denotes the normalized structural distance between v_i and v_j, $d_{\mathrm{A}}(v_i, v_j)$ represents the distance between their attributes, and $\alpha \in [0, 1]$ is a weighting parameter. We consider $\mathbf{D} = [d(v_i, v_j)]_{1 \leq i,j \leq n}$ as the symmetric distance matrix representing the distances between all nodes. While in the initial article, d_A represents a distance for textual attributes, traffic-related data require other choices of metrics, which will be discussed in subsection 2.3.

2.1 k Fréchet means

The k-means algorithm with Fréchet means aims to partition a set of n data points $\{x_1, \ldots, x_n\}$ into k clusters $\{C_1, \ldots, C_k\}$, while selecting k clusters centers $\{m_1, \ldots, m_k\} \in \{x_1, \ldots, x_n\}$ that minimize the total within-cluster dissimilarity.

$$\min_{\{C_i\}_{i=1}^k, \{m_i\}_{i=1}^k} \sum_{i=1}^{k} \sum_{x \in C_i} d^2(x, m_i)\mu(x)$$

where d is a distance or dissimilarity measure, m_i is a Fréchet mean of cluster C_i, i.e., $m_i \in \arg\min_{y \in C_i} \sum_{x \in C_i} d^2(x, y)\mu(x)$ and $C_i \subset \{x_1, \ldots, x_n\}$ is the set of points assigned to the i-th cluster.

Since an attributed graph can be described by the distance matrix $\mathbf{D}$ between its nodes, this type of methods can be adapted to node-attributed graph clustering. We use the k-means algorithm with this distance in Algorithm 1.

Algorithm 1 k Fréchet means clustering for node-attributed graph clustering

Input: An attributed weighted graph $G = (V, E, A)$ characterized by its pair-wise distance matrix $\mathbf{D} = [d(v_i, v_j)]_{1 \leq i,j \leq n}$, k the number of desired subsets, initial clusters centers $m_1, \ldots, m_k$ and a maximum number of iterations
repeat
  for each node $v_i \in V$ **do**
    Assign v_i to the nearest cluster center, i.e., $\arg\min_{1 \leq j \leq k} d(m_j, v_i)\mu(v_i)$
  end for
  for each cluster C_j **do**
    Update the cluster center: $m_j = \arg\min_{x \in C_j} \sum_{v \in C_j} d^2(v, x)\mu(v)$
  end for
until Convergence (no change in cluster centers) or maximum number of iterations is reached.
return Subsets $\{C_1, \ldots, C_k\}$ and their centers $\{m_1, \ldots, m_k\}$

2.2 Partitioning through Optimal Transport

Gromov-Wasserstein metric A graph G can be modeled as a couple $(\mathbf{C}, \boldsymbol{\mu})$, with $\mathbf{C} \in \mathbb{R}^{n \times n}$ representing the connectivity between nodes (adjacency matrix or shortest path distance for example) and $\boldsymbol{\mu} \in \Sigma_n$, the probability simplex. While classical Optimal Transport (OT) aims to compare probability distributions with supports in the same metric space, Peyré, Cuturi, and Solomon (2016) propose a method for comparing two distance matrices that do not share the same ground space, using the Gromov-Wasserstein (GW) discrepancy (Mémoli

2011). This approach seeks to find a matching between these distributions that is as close as possible to an isometry, thus allowing to define a notion of distance between two graphs. Considering two graphs, $G_1 = (\mathbf{C}^{(1)}, \boldsymbol{\mu}^{(1)})$ and $G_2 = (\mathbf{C}^{(2)}, \boldsymbol{\mu}^{(2)})$, with respectively n and k nodes, the Gromow-Wasserstein discrepancy is defined as:

$$GW(\mathbf{C}^{(1)}, \boldsymbol{\mu}^{(1)}, \mathbf{C}^{(2)}, \boldsymbol{\mu}^{(2)}) = \min_{\substack{\mathbf{T}\mathbf{1}_k = \boldsymbol{\mu}^{(1)} \\ \mathbf{T}^T\mathbf{1}_n = \boldsymbol{\mu}^{(2)}}} \sum_{i,j,l,m} |C^{(1)}_{ij} - C^{(2)}_{lm}|^2 T_{il} T_{jm}$$

with $\mathbf{1}_n$ is a column vector of size n where all entries are equal to 1 and $\mathbf{T} \in \mathbb{R}_+^{n \times k}$ the optimal transport plan, representing the probabilistic matching of nodes.

Semi-relaxed Gromov-Wasserstein divergence Vincent-Cuaz et al. (2022) consider that the OT core idea of mass conservation could be detrimental for some unsupervised tasks and propose to relax the second marginal. They define a semi-relaxed GW divergence:

$$srGW(\mathbf{C}^{(1)}, \boldsymbol{\mu}^{(1)}, \mathbf{C}^{(2)}) = \min_{\boldsymbol{\mu}^{(2)} \in \Sigma_k} GW(\mathbf{C}^{(1)}, \boldsymbol{\mu}^{(1)}, \mathbf{C}^{(2)}, \boldsymbol{\mu}^{(2)}) = \min_{\mathbf{T}\mathbf{1}_k = \boldsymbol{\mu}^{(1)}} \sum_{i,j,l,m} |C^{(1)}_{ij} - C^{(2)}_{lm}|^2 T_{il} T_{jm}$$

Application to graph partitioning Graph partitioning raises the important question of selecting an appropriate target structure for the graph. Xu, Luo, and Carin (2019) propose using a disconnected graph with k isolated, self-connected nodes as the target structure. The central idea is to treat each partition, or subgraph, as a super-node, with the ideal partitioning corresponding to an entirely disconnected target graph. The resulting OT plan provides a soft assignment of each node to the super-nodes, thus defining the partition. Whereas the GW metric requires the target distribution to be fixed, thereby determining the size of the classes, the semi-relaxed GW divergence introduces a relaxation of the mass constraints on the target marginals. Because $\boldsymbol{\mu}^{(2)}$ is optimized within the simplex Σ_k, it may exhibit sparsity. The mass might be transported to fewer than k nodes, alleviating the traditional challenge of selecting k in unsupervised clustering. Vincent-Cuaz et al. (2022) previously evaluate the semi-relaxed GW divergence for graph partitioning but focus exclusively on non-attributed graphs. In their work, the adjacency matrix or normalized Laplacian are used as $\mathbf{C}$ for partitioning. We propose extending this approach by using $\mathbf{C} = \mathbf{D}$, the distance matrix that combines topological and attribute-based distances (Algorithm 2).

Algorithm 2 semi-relaxed Gromov-Wasserstein for node-attributed graph clustering

Input: An attributed weighted graph $G = (V, E, A)$ characterized by its pair-wise distance matrix $\mathbf{D} = [d(v_i, v_j)]_{1 \leq i,j \leq n}$, k the maximum number of desired subsets and an initial transportation plan
Create the target matrix $\mathbf{M} = \mathbf{1}_{k \times k} - I_k$
Get optimal transport plan $\mathbf{T}^*$ from $srGW(\mathbf{D}, \boldsymbol{\mu}, \mathbf{M})$ with conditional gradient algorithm
Get the subsets $C_i = \{j \mid T^*_{ji} \geq T^*_{jl}, \forall l \in \{1, \ldots, k\}\}$
return Subsets $\{C_1, \ldots, C_k\}$ where some C_i may be empty.

2.3 Considered Distances for Traffic Attributed Graphs

Graph Distance For the structural distance d_S, the physical distance is considered, represented by the length of the shortest path between two nodes, normalized to be combined with the attribute distances: $\tilde{d}_S(v_i, v_j) = \frac{d_S(v_i,v_j)-\min(d_S)}{\max(d_S)-\min(d_S)}$.

Attributes Distance Traffic data may exhibit slight time shift in their peaks across different geographical zones, but the primary objective is to detect similarity in shape. Therefore, we employed Dynamic Time Warping (or DTW), denoted $d_{\text{dtw}}(f_i, f_j)$, where f_i and f_j are the functional data associated to nodes i and j, respectively. Histograms data are compared using the Wasserstein 1-distance, denoted $W_1(\mathbf{h_i}, \mathbf{h_j})$, where $\mathbf{h_i}$ (resp. $\mathbf{h_j}$) is the discrete distribution associated to node i (resp. node j). To enable comparison, the attribute distances are also rescaled using min-max normalization. Ultimately, the attributes distance are combined: $d_A(v_i, v_j) = \beta \cdot \tilde{d}_{\text{dtw}}(f_i, f_j) + (1-\beta) \cdot \tilde{W}_1(\mathbf{h_i}, \mathbf{h_j})$, $\beta \in [0, 1]$ a weighting parameter.

3 Simulation and Real-World Data Description

The methods are currently being tested on synthetic data and real-world data. The proposed simulation framework simulate both a graph structure and a set of functional attributes and histograms associated with it, while the real-world data provide a road network with traffic patterns and speed histograms.

3.1 Synthetic data

Functional Data Generation Basis splines (or B-splines) are used to generate traffic patterns. For each group $j \in \{1, \ldots, k\}$, different coefficients are independently sampled and associated with each basis function. These coefficients are then perturbed to produce n_j noisy versions.

Histograms Generation A Dirichlet distribution with randomly drawn concentration parameters is used to generate k probability vectors representing speeds. For each vector, multiple variants are sampled from another Dirichlet distribution, with parameters adjusted to control the proximity of the variants to the original vector.

Graph Generation The stochastic block model (SBM) simulates the road network, with a block matrix encoding edge probabilities between groups. In the simulation, a random $k \times k$ matrix is generated, and diagonal elements are modified to reflect higher intra-group connection density.

Testing Framework: Scenarios and Parameters Testing methods involve varying the relationship between node attributes and SBM groups. The first and simplest scenario is to assign attributes consistently with the SBM groups. A second version evaluate a hierarchical attribution, where two versions of attributes coexist within the same SBM group or if two groups share the same attributes, creating different levels of complexity. Monte Carlo simulations will evaluate sensitivity to attribute noise and the impact of the group structure strength.

3.2 Traffic Data

Floating Car Data (FCD) refer to information collected from vehicles equipped with GPS or other tracking technologies. These data provide information into vehicle speeds and the number of vehicles in motion at specific times. The FCD used in this study pertains to the city of Châteaubourg (Brittany), a small city that experiences recurrent congestion issues. This city was selected due to an ongoing citizen science project, that aims to measure traffic to better understand car mobility in low-density areas and support informed public policy decisions. Since the ground truth for the partitioning is unknown, the resulting partitions of road network can also be interpreted and validated using socio-economic indicators, such as household composition or average income.

4 Perspectives

Another way to adapt the semi-relaxed GW divergence to attributed graphs is already detailed by Vincent-Cuaz et al. (2022), using the Fused Gromov-Wasserstein method (Vayer et al. 2019). This method was initially developed for optimal transport on structured data (such as attributed graphs). However, as it is not originally designed for graph partitioning, its adaptation could address current limitations related to attribute types and propose a more refined approach to account for both structure and attributes in partitioning.

Acknowledgement This work was conducted within the France 2030 program, Centre Henri Lebesgue ANR-11-LABX-0020-01.

References

[1] Combe, David et al. “Combining Relations and Text in Scientific Network Clustering”. In: *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2012.

[2] Fortunato, Santo. “Community Detection in Graphs”. In: *Physics Reports* 3 (2010).

[3] Mémoli, Facundo. “Gromov–Wasserstein Distances and the Metric Approach to Object Matching”. In: *Foundations of Computational Mathematics* (2011).

[4] Peyré, Gabriel, Cuturi, Marco, and Solomon, Justin. “Gromov-Wasserstein Averaging of Kernel and Distance Matrices”. In: *Proc. 33rd International Conference on Machine Learning*. 2016.

[5] Vayer, Titouan et al. “Optimal Transport for structured data with application on graphs”. In: *Proceedings of the 36th International Conference on Machine Learning*. 2019.

[6] Vincent-Cuaz, Cédric et al. *Semi-relaxed Gromov-Wasserstein divergence with applications on graphs*. 2022. arXiv: 2110.02753.

[7] Xu, Hongteng, Luo, Dixin, and Carin, Lawrence. *Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching*. 2019. arXiv: 1905.07645.

The Deep Zero-Inflated Latent Position Block Model for the Clustering of Nodes in Graphs

S. NIANG[1] & C. BOUVEYRON[1] & M. CORNELI[1,2] & P. LATOUCHE[3,4]

[1] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France*
[2] *Université Côte d'Azur, Laboratoire CEPAM, Nice, France*
[3] *Université Clermont Auvergne, CNRS, Laboratoire LMBP, Aubière, France*
[4] *Institut Universitaire de France (IUF)*

Résumé. L'augmentation des capacités de stockage a entraîné une explosion des données, rendant les réseaux essentiels pour modéliser les relations entre objets (nœuds). Ces réseaux, souvent complexes, nécessitent des méthodes efficaces de clustering et de visualisation pour en résumer l'information. Le deep latent position block model (Deep-LPBM), conçu pour les réseaux binaires, combine clustering par blocs et représentation latente continue pour visualiser les nœuds. Nous proposons ici une extension, le deep zero-inflated latent posistion block model (Deep-ZLPBM), adapté aux réseaux non binaires, où la matrice d'adjacence peut prende des valeurs entières. Ce modèle repose sur un auto-encodeur variationnel intégrant un réseau de convolution de graphes (GCN) et un décodeur utilisant une loi ZIP (zero-inflated Poisson). L'inférence se concentre sur la maximisation de la vraisemblance marginale, et l'optimisation s'effectue par descente de gradient.

Mots-clés. Clustering de nœuds, Auto-encodeur variationnel pour graphe, Modélisation par blocs, Visualisation de graphes, Zero-inflatted Poisson

Abstract. The evolution in storage capacities has led to a data explosion, making networks essential for modeling relationships between objects (nodes). These complex networks require effective clustering and visualization methods to summarize and interpret their information. The deep latent position block model (Deep-LPBM), designed for binary networks, combines partial block-based clustering and continuous latent representation to visualize nodes. Here, we propose an extension, the deep zero-inflated latent position block model (Deep-ZLPBM), designed for non-binary networks, where the entries of the adjacency matrix can take integer values. This model is based on a deep variational autoencoder that integrates a graph convolutional network (GCN) and a decoder leveraging a zero-inflated Poisson (ZIP) distribution. Inference relies on the maximization of the marginal likelihood, and optimization is performed using stochastic gradient descent.

Keywords. Nodes clustering, Graph variational autoencoder, Block modelling, Graph visualization, Zero-inflated Poisson

1 Introduction

Networks represent connections among entities, such as individuals, organizations, or biological elements. Nodes are entities and edges characterize their relationships. Social

networks are a common example, with nodes representing individuals and edges representing friendship or collaborations. As network data grows in scale and complexity, understanding these interconnections becomes increasingly challenging. The interdependence between nodes creates heterogeneous connectivity patterns, requiring models that balance complexity and interpretability. Clustering nodes [1] is a key approach in network analysis, offering a high-level summary of complex networks. While many studies focus on binary networks, where connections are either present or absent, valued networks are equally important. These networks capture rich details, such as interaction frequencies or relationship strengths. This paper explores model-based clustering for valued networks, where connections are modeled probabilistically.

Historical models for binary networks [2, 3] relied on simple probabilistic assumptions, like fixing edge probability constant. Yet, real-world networks exhibit characteristics that defy these simplified approaches. Pioneering works allowed for the introduction of advanced models and among them, latent variable models (LVMs) have gained significant attention. LVMs assume that interconnections between nodes depend on unobserved latent variables. The earliest and most popular LVMs for node clustering include the latent position cluster model (LPCM) [4] and the stochastic block model(SBM) [5]. The SBM efficiently clusters nodes into groups and supposes that interactions between two nodes only depend on the clusters they belong to, but does not allow any direct latent representation (visualisation) of the underlying structure of the graph. The LPCM allows to have a direct latent visualisation of the network but this clustering approach only excels at community detection and lacks the flexibility to capture more diverse connectivity patterns. Recent developments in this area include [6], which uses deep variational inference procedure and extends the LPCM to textual data analysis.

One limitation of both SBMs and LPCMs is the assumption they make that each node belongs to a single group. In reality, individuals often belong to multiple overlapping groups, reflecting the multiple roles they play within several groups. Many works have been proposed to account for such partial memberships [7, 8].
Recently, in [9], the deep latent position block model (Deep-LPBM) was proposed to combine the strengths of SBM as well as LPCM, and to deal with partial memberships. Deep-LPBM enables both latent visualization and partial membership block modeling, providing a more versatile clustering framework than traditional positional clustering methods.

In this work, we introduce the deep zero inflated latent position block model (Deep-ZLPBM). This framework is built within the class of latent position block models, extending [9] to the analysis of valued networks. Several works have also attempted to extend binary networks methods to integer-valued networks. For example in [10], the degree-corrected stochastic block model (DCSBM) is proposed. DCSBM extends the traditional SBM to valued networks by mainly replacing the Bernoulli distribution by a Poisson [11]. Many other works have adopted the same strategy [12, 13].

2 Generative models

In this work, we suppose that the graph is undirected and identified by its adjacency matrix A, a square $N \times N$ symmetric matrix such that its element $A_{ij} = A_{ji} \in \mathbb{N}$.

2.1 Generative model for the graph

We assume that the N nodes belong to Q clusters in a non exclusive manner and that each node node i belongs to cluster q with a probability η_{iq} as a function of the latent variable $Z_i \overset{\text{iid}}{\sim} \mathcal{N}_d(0, \mathbb{I}_{Q-1})$. The set of node embeddings is denoted by $Z := \{Z_i\}_i$ in the rest of the paper.

To link the latent representation of the nodes Z_i with partial membership probabilities $\eta_i \in \Delta_Q := \{x \in [0,1]^Q; \sum_{q=1}^Q x_q = 1\}$, we rely on the bijective softmax transformation $h : \mathbb{R}^{Q-1} \to \Delta_Q$ such that:

$$\eta_{iq} := h(Z_i)_Q := \begin{cases} \frac{\exp(Z_{iq})}{1+\sum_{r=1}^{Q-1}\exp(Z_{ir})} & \text{if} \quad q \neq Q, \\ \frac{1}{1+\sum_{r=1}^{Q-1}\exp(Z_{ir})} & \text{if} \quad q = Q, \end{cases} \tag{1}$$

and we denote $\eta = (\eta_1, \ldots, \eta_N)^\intercal$. The mapping h aims at encoding Z into cluster partial membership probabilities. We assume that the probability of connection between two nodes follows a zero-inflated Poisson distribution [14] with parameters depending on η such that:

$$\begin{aligned} A_{ij}|Z_i, Z_j &\sim \text{ZIP}(\eta_i^\intercal \Pi \eta_j, \eta_i^\intercal \Lambda \eta_j), \\ p(A, Z|\Pi, \Lambda) &= \left(\prod_{i=1}^N p(Z_i)\right) p(A|Z, \Pi, \Lambda), \end{aligned} \tag{2}$$

where, adapting the simplified notation $\prod_{i<j} = \prod_{i=1}^{N-1}\prod_{j=i+1}^{N}$, we have

$$\begin{aligned} p(A|Z, \Lambda, \Pi) &= \prod_{i<j} p(A_{ij}|Z, \Pi, \Lambda) \\ &= \prod_{\substack{i<j \\ A_{i,j}=0}} \left[1 - \eta_i^\intercal \Pi \eta_j (1 - \exp\left(-\eta_i^\intercal \Lambda \eta_j\right))\right] \prod_{\substack{i<j \\ A_{i,j}>0}} \left[\eta_i^\intercal \Pi \eta_j \, \frac{(\eta_i^\intercal \Lambda \eta_j)^{A_{ij}}}{A_{ij}!} \exp(-\eta_i^\intercal \Lambda \eta_j)\right] \end{aligned} \tag{3}$$

where the $Q \times Q$ matrix $\Pi = (\Pi_{qr})_{1 \leq q,r \leq Q}$ refers to the connectivity matrix whose entry (q, r) is the probability that a node in block q is connected to a node in block r (i.e. the corresponding value of the adjacency matrix is not zero). The parameter Λ is also a $Q \times Q$ matrix whose entry (q, r) refers to the expectation of A_{ij} with i and j respectively belonging to clusters q and r.

.

3 Variational inference and optimisation

The marginal log-likelihood of the graph is given by

$$\log p(A|\Pi, \Lambda) = \log \int_Z p(A, Z|\Pi, \Lambda) dZ.$$

Unfortunately, this quantity cannot be computed analytically. Consequently, we propose to rely on a variational inference strategy for approximation purposes. For any distribution $q(\cdot)$ for the latent variables in Z, in force of the Jensen's inequality, the following

decomposition holds:

$$\log p(A|\Pi,\Lambda) = \log \int_Z \frac{p(A,Z|\Pi,\Lambda)}{q(Z)} q(Z) dZ = \log \mathbb{E}_{q(Z)}\left[\frac{p(A,Z|\Pi)}{q(Z)}\right]$$
$$\geq \mathbb{E}_{q(Z)}\left[\log \frac{p(A,Z|\Pi)}{q(Z)}\right] := \mathcal{L}(\Pi,\Lambda,q(\cdot)).$$

In this paper, we refer to $\mathcal{L}(\Pi,\Lambda,q(\cdot))$ as the evidence lower bound (ELBO). Furthermore, the exact difference between $\log p(A|\Pi,\Lambda)$ and $\mathcal{L}(\Pi,\Lambda,q(\cdot))$ is

$$\mathrm{KL}(q(\cdot)||p(Z|A,\Pi)) = \log p(A|\Pi,\Lambda) - \mathcal{L}(\Pi,\Lambda,q(\cdot)), \tag{4}$$

where the Kullback-Leibler (KL) divergence between q and the posterior distribution $p(Z|A,\Pi)$ is always non-negative, indicating that the ELBO is a lower bound of the marginal log-likelihood. Since the marginal log-likelihood does not depend on $q(\cdot)$, maximizing the ELBO with respect to $q(\cdot)$ is equivalent to minimizing the Kullback-Leibler divergence between $q(\cdot)$ and the posterior distribution. This minimisation is satisfied when the variational $q(\cdot)$ is equal to the true posterior distribution $p(Z|A,\Pi)$ which is not tractable here, so we adopt a mean-field approximation and other hypotheses to make the ELBO tractable:

$$q(Z) := q(Z|A) = \prod_{i=1}^{N} q_\phi(Z_i|A) = \prod_{i=1}^{N} \mathcal{N}(Z_i; \mu_\phi(\bar{A})_i, \sigma^2_\phi(\bar{A})_i \mathbb{I}_{Q-1}), \tag{5}$$

with $\left[\mu_\phi(\bar{A}), \sigma_\phi(\bar{A})\right] = f_\phi(\bar{A})$, where $f_\phi : \mathbb{R}^{N\times N} \to \mathbb{R}^Q$ is a graph convolutional network (GCN, [15]) parameterized by ϕ that maps the normalized adjacency matrix $\bar{A} = \tilde{D}^{-\frac{1}{2}}(A + I_N)\tilde{D}^{-\frac{1}{2}}$ to the vector of the variational means and the log standard deviations. Here, $\tilde{D}$ is the diagonal matrix representing the degree of nodes with respect to the matrix $(A+I_N)$, defined as $\tilde{D}_{ii} = 1 + \sum_{j=1}^{N} A_{ij}$. In Equation (5), $\mu_\phi(\bar{A})_i$ denotes the i-th row of $\bar{\mu}_\phi(A)$, and $\bar{\sigma}^2_\phi(A)_i$ denotes the i-th element of $\sigma^2_\phi(A)$. Thus, the ELBO can be decomposed as follows:

$$\begin{aligned}
\mathcal{L}(q,\Lambda,\Pi) &= \mathbb{E}_{q(\cdot)}\left[\log \frac{p(A,Z|\Pi,\Lambda)}{q(Z|A)}\right] \\
&= \mathbb{E}_{q(\cdot)}\left[\log \frac{p(A|Z,\Pi,\Lambda)p(Z)}{q(Z|A)}\right] \\
&= \underbrace{\mathbb{E}_{q(\cdot)}\left[\log p(A|Z,\Pi,\Lambda)\right]}_{\text{Reconstruction term}} - \underbrace{\mathrm{KL}(q(Z|A)||p(Z))}_{\text{Regularisation term}}.
\end{aligned}$$

Finally:

$$\begin{aligned}
\mathcal{L}(q,\Pi,\Lambda) = \mathbb{E}_{q(\cdot)}\Bigg[&\sum_{\substack{i<j \\ A_{i,j}=0}} \log(1 - \eta_i^\intercal \Pi \eta_j (1 - \exp(-\eta_i^\intercal \Lambda \eta_j))) \\
&+ \sum_{\substack{i<j \\ A_{i,j}>0}} \left[\log(\eta_i^\intercal \Pi \eta_j) + A_{ij}\log(\eta_i^\intercal \Lambda \eta_j) - \log(A_{ij}!) - \eta_i^\intercal \Lambda \eta_j\right]\Bigg] - \\
&\sum_i \left[(Q-1)\log\sigma_\phi(\bar{A})_i - \frac{Q-1}{2} + \frac{1}{2}\left|\left|\mu_{\phi,\xi}(\bar{A})_i\right|\right|_2^2 + \frac{Q-1}{2}\sigma_\phi(\bar{A})_i^2\right].
\end{aligned} \tag{6}$$

The regularization term, which involves the KL divergence between q and the prior distribution, can be computed exactly. The main challenge lies in calculating the reconstruction term. To address this and optimize the ELBO, we introduce a Monte Carlo gradient descent algorithm, detailed in the next section.

3.1 Monte Carlo variational optimisation

The model and variational parameters cannot be updated with analytical formulas because of the integral involving the variational distribution $q(\cdot)$ in the ELBO. In this section, we aim at deriving estimates $\tilde{\mathcal{L}}(\cdot)$ of the ELBO $\mathcal{L}(\cdot)$ to perform Monte Carlo gradient descent:

$$\begin{aligned}\tilde{\mathcal{L}}(q,\Pi,\Lambda) = \frac{1}{S}\sum_{s=1}^{S}\Bigg[&\sum_{\substack{i<j\\A_{i,j}=0}} \log\Big(1-\eta_i^{\intercal(s)}\Pi\eta_j^{(s)}\big(1-\exp\big(-\eta_i^{\intercal(s)}\Lambda\eta_j^{(s)}\big)\big)\Big)\\ &+\sum_{\substack{i<j\\A_{ij}>0}}\Big[\log\Big(\eta_i^{\intercal(s)}\Pi\eta_j^{(s)}\Big)+A_{ij}\log\Big(\eta_i^{\intercal(s)}\Lambda\eta_j^{(s)}\Big)-\log(A_{ij}!)-\eta_i^{\intercal(s)}\Lambda\eta_j^{(s)}\Big]\Bigg]\\ &-\sum_i\left[(Q-1)\log\sigma_\phi(\bar{A})_i-\frac{Q-1}{2}+\frac{1}{2}\left\|\mu_\phi(\bar{A})_i\right\|_2^2+\frac{Q-1}{2}\sigma_\phi(\bar{A})_i^2\right],\end{aligned} \tag{7}$$

where $\eta_i^{(s)} = h(Z_i^{(s)})$ and $Z_i^{(s)} \sim q(\cdot)$. This usual (naïve) Monte Carlo gradient estimator exhibits very high variance and is then impractical [16]. To deal with this problem, as done in [17], we adopt the reparameterization trick. In particular, if $\epsilon^{(s)} \sim \mathcal{N}(0, \mathbb{I}_{Q-1})$, then $Z_i^{(s)} = \mu_\phi(\bar{A})_i \ + \sigma_\phi(\bar{A})_i \ \epsilon^{(s)} \sim \mathcal{N}(\mu_\phi(\bar{A})_i, \sigma_\phi(\bar{A})_i^2\mathbb{I}_{Q-1}) = q(\cdot|A)$. To simplify the gradient descent, we map the constrained values $(\Pi_{qr})_{q,r}$ (resp $(\Lambda_{qr})_{q,r}$) from the interval $]0,1[$ (resp $]0,+\infty[$) to the unconstrained set $\mathbb{R}$ using the bijective function g (resp f) defined as:

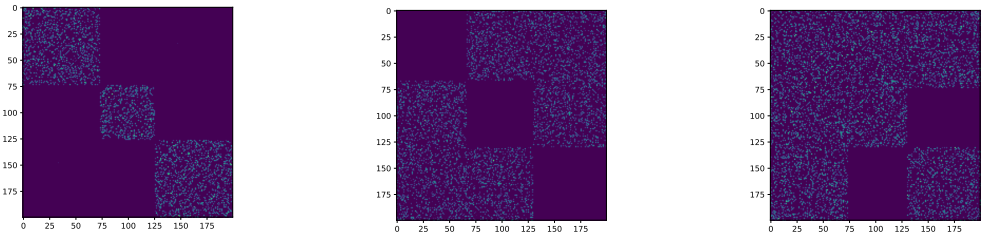
$$g: \begin{cases}\mathbb{R}\to]0,1[\\ x\mapsto 0.5+\frac{1}{\pi}\arctan(x)\end{cases} \qquad \text{resp} \quad f: \begin{cases}\mathbb{R}\to]0,\infty[\\ x\mapsto \exp(x)\end{cases}$$

.

4 Evaluation on synthetic datasets

To evaluate the ability of our method to represent diverse network topologies, we tested it on three different network structures, each consisting of 200 nodes spread into $Q = 3$ clusters. For the connectivity matrix Π, we considered three different network structures each reflecting a kind of social interaction: i) the **community structure**, where we assume that nodes in the same group have a high probability of connection indicated by β and that nodes in different groups have a low probability of connection indicated by δ, ii) the **disassortative structure** where we assume that nodes in different clusters have a high probability of connection denoted by β, and that nodes in the same cluster have a low probability of connection denoted by δ, iii) a **hub structure** where we assume that one of the clusters is highly connected to all the clusters, with a probability β and that the other clusters are communities. For the expectation matrix Λ, we set for all $q < r$,

$\Lambda_{qr} = \alpha$ and $\Lambda_{qq} = \sigma$.
To generate the adjacency A, we fix the labels and one-hot them to build κ such that $\kappa_{iq} = 1$ if node i is in cluster q and 0 otherwise. We then generate $M \sim Bernoulli(\cdot|1; \kappa^\intercal \Pi \kappa)$ and $N \sim Poisson(\cdot|\kappa^\intercal \Lambda \kappa)$ and set $A = M \times N$ (Hadamard product). In Figure 1, we plot some adjacency matrices, reorganized by blocs, of the three different structures presented above with different connectivity levels $\beta = 0.2, \delta = 0.01, \alpha = 5$ and $\sigma = 0.03$.
The adjusted rand index (ARI, [18]) serves as our primary measure of clustering accuracy, reflecting the similarity between the true and inferred node partitions. An ARI of 0 suggests clustering is no better than random, while an ARI close to 1 indicates alignment with true node labels up to label switching. In Table 1, we show the performance of Deep-ZLPBM on three different connectivity levels.



Community Disassortative Hub

Figure 1: Examples of adjacency matrices corresponding to the three structures and connectivity levels $\beta = 0.2, \delta = 0.01, \alpha = 5$ and $\sigma = 0.03$ and reorganised by cluster.

Connectivity levels	ARI		
	Communities	Disassortative	Hub
$\beta = 0.3, \delta = 0.01, \alpha = 5, \sigma = 0.03$	1 ± 0.00	1 ± 0.00	0.98 ± 0.01
$\beta = 0.2, \delta = 0.01, \alpha = 5, \sigma = 0.03$	1 ± 0.00	1 ± 0.00	0.8 ± 0.09
$\beta = 0.1, \delta = 0.01, \alpha = 5, \sigma = 0.03$	0.96 ± 0.08	0.48 ± 0.23	0.2 ± 0.13

Table 1: ARI of Deep–ZLPBM on different connectivity levels.

When the connectivity levels are high ($\beta \in \{0.3, 0.2\}$), in both communities and disassortative structures, Deep-ZLPBM efficiently recovers the true nodes partition (ARI $= 1$). In the Hub configuration, it is harder for our method to recover all the clusters (ARI $\in \{0.98, 0.8\}$). For $\beta = 0.1$, the problem becomes much harder. Deep-ZLPBM performs well in the community structure but struggle with hub and disassortative.

5 Conclusion

This paper extended the Deep-LPBM, a new methodology combining a block model with latent position model adapted to valued networks. A preliminary study with simulated data illustrated the capacity of our methodology to clusters nodes in valued networks. This is an on going work and an extensive benchmark will be provided if the paper were to be accepted.

References

[1] S. E. Schaeffer, "Graph clustering," *Computer science review*, vol. 1, no. 1, pp. 27–64, 2007.

[2] O. Erdős and A. Renyi, "On random graphs," *Publ. Math*, vol. 6, pp. 290–297, 1959.

[3] P. W. Holland and S. Leinhardt, "An exponential family of probability distributions for directed graphs," *Journal of the american Statistical association*, vol. 76, no. 373, pp. 33–50, 1981.

[4] P. D. Hoff, A. E. Raftery, and M. S. Handcock, "Latent space approaches to social network analysis," *Journal of the american Statistical association*, vol. 97, no. 460, pp. 1090–1098, 2002.

[5] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.

[6] R. Boutin, P. Latouche, and C. Bouveyron, "The deep latent position topic model for clustering and representation of networks with textual edges," *arXiv preprint arXiv:2304.08242*, 2023.

[7] K. A. Heller, S. Williamson, and Z. Ghahramani, "Statistical models for partial membership," in *Proceedings of the 25th International Conference on Machine learning*, pp. 392–399, 2008.

[8] E. M. Airoldi, D. M. Blei, S. E. Fienberg, E. P. Xing, and T. Jaakkola, "Mixed membership stochastic block models for relational data with application to protein-protein interactions," in *Proceedings of the international biometrics society annual meeting*, vol. 15, p. 1, 2006.

[9] R. Boutin, P. Latouche, and C. Bouveyron, "The deep latent position block model for the block clustering and latent representation of networks," *arXiv preprint arXiv:2412.01302*, 2024.

[10] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 83, no. 1, p. 016107, 2011.

[11] P. C. Consul and G. C. Jain, "A generalization of the poisson distribution," *Technometrics*, vol. 15, no. 4, pp. 791–799, 1973.

[12] T. Herlau, M. N. Schmidt, and M. Mørup, "Infinite-degree-corrected stochastic block model," *Physical review E*, vol. 90, no. 3, p. 032819, 2014.

[13] C. Aicher, A. Z. Jacobs, and A. Clauset, "Learning latent block structure in weighted networks," *Journal of Complex Networks*, vol. 3, no. 2, pp. 221–248, 2015.

[14] D. Lambert, "Zero-inflated poisson regression, with an application to defects in manufacturing," *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.

[15] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[16] J. Paisley, D. Blei, and M. Jordan, "Variational bayesian inference with stochastic search," *arXiv preprint arXiv:1206.6430*, 2012.

[17] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[18] K. Y. YEUNG and W. L. RUZZO, "Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

Méthode intégrée d'inférence de réseaux et de clustering des variables sur des données continues: le modèle Normal-Block

Jeanne Tous*†[1] and Julien Chiquet[1]

[1]Mathématiques et Informatique Appliquées – AgroParisTech, Université Paris-Saclay, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UMR0518 – France

Résumé

On considère le problème d'estimation des paramètres d'un modèle graphique gaussien pour un grand nombre de variables. Ces modèles permettent de décrire des structures de dépendance entre des variables qui peuvent être des gènes, des personnes, des espèces, etc. Lorsque le nombre de variables considérées augmente, l'inférence des paramètres du modèle et l'interprétation du réseau d'association qui en résulte deviennent plus complexes. Nous proposons le modèle Normal-Block qui permet de réduire la dimension du réseau afin de faciliter son inférence et son analyse. Pour ce faire, le modèle suppose un clustering sur les variables et considère un réseau à l'échelle des clusters plutôt qu'à celle des variables individuelles. Nous utilisons la méthode Graphical-Lasso pour intégrer une pénalité sur les arêtes du réseau afin de limiter la détection de faux positifs dans les associations et proposons une version zéro-inflatée de notre modèle pour mieux l'adapter aux données réelles. Pour l'inférence des paramètres, nous proposons une méthode heuristique directe ainsi qu'une méthode intégrée, plus rigoureuse, fondée sur un algorithme d'expectation-maximization. Cette seconde méthode infère simultanément le clustering et le réseau d'association entre clusters. Les tests sur données simulées montrent de bonnes performance des méthodes d'inférence à la fois pour l'évaluation du clustering et celle du réseau.

Mots-Clés: Modèles de mélange, modèles graphiques gaussiens, clustering, inférence variationnelle.

*Intervenant

†Auteur correspondant: jeanne.tous@inrae.fr

Culture statistique (Session groupe Enseignement)

PLUS DE CULTURE STATISTIQUE = PLUS DE CITOYENNETÉ ACTIVE : DONNÉES ET MÉTHODES

Simona Cafieri

Istat, Italie, cafieri@istat.it

Résumé
Au niveau international, l'importance de la statistique publique est désormais reconnue : elle est indispensable pour que chaque citoyen puisse connaître et reconnaître de manière critique le contexte dans lequel il vit et être capable de prendre des décisions, même en situation d'incertitude, sur la base de données concrètes et partagées.

Les instituts nationaux de statistique, animés par la conviction qu'une société où les décisions sont prises sur la base de données objectives est une société où règne la démocratie, se sont efforcés de rendre l'accès aux données officielles gratuit et facilement accessible à leurs citoyens. Une réponse immédiate donc, mais non sans écueils car le libre accès à l'information statistique est un moyen d'assurer le progrès d'une société démocratique, mais il perd de son efficacité si les citoyens ne sont pas en mesure de lire, d'interpréter et d'analyser les données. Donc.il s'agissait d'un nouveau défi:non seulement diffuser les données, mais les promouvoir, apprendre à les connaître, les faire comprendre. pour qu'elles soient utiles à la communauté, un bien qui a un fort impact potentiel sur la vie de chacun.

C'est pourquoi il peut être très utile d'acquérir des compétences en matière de statistiques et de lecture de données dès l'école. Non pas tant pour devenir des statisticiens à l'âge adulte, mais pour apprendre à lire et à décoder correctement le monde qui nous entoure.

Parmi les différents projets qui visent à développer la culture statistique dans les écoles, ce travail entend illustrer un parcours didactique interdisciplinaire innovant visant à promouvoir et développer les principes d'une citoyenneté active et consciente dans les écoles, à travers des activités de recherche, l'utilisation de données ouvertes, l'utilisation des technologies de l'information et le suivi civique des financements publics européens et nationaux.

Ce projet, destiné aux collèges et lycées, permet de développer des compétences numériques, statistiques et d'éducation civique, ainsi que d'autres compétences transversales telles que le développement du sens critique et la résolution de problèmes, le travail en équipe, les compétences interpersonnelles et de communication, en les intégrant aux contenus des matières d'étude ordinaires, afin d'aider les élèves à connaître et à communiquer, à l'aide de techniques journalistiques, la manière dont les politiques publiques, et en particulier les politiques de cohésion, interviennent dans les lieux où ils vivent.

L'enseignement est basé sur des projets, combinant des moments d'apprentissage asynchrones typiques des MOOCs (Massive Online Open Courses) avec des activités de facilitation en présence dirigées par les enseignants eux-mêmes (préalablement formés ad hoc), des travaux de groupe et des interactions en ligne avec l'équipe du projet.

Les élèves découvriront donc en classe ce à quoi servent les données avec l'aide de leurs enseignants, dans un voyage informatif intéressant dans le monde des chiffres, qui les aidera à développer leur capacité à comprendre les évolutions sociales, culturelles et environnementales grâce aux données fournies par les statistiques officielles. De plus, ils pourront participer à une initiative pratique, en

s'essayant à la création d'un produit de communication et en mettant à l'épreuve leurs capacités d'analyse et leur créativité.

Mots-clés. numératie, culture statistique, surveillance civique, citoyenneté active

MORE STATISTICAL LITERACY = MORE ACTIVE CITIZENSHIP: DATA AND METHODS

Abstract

At international level, the importance of official statistics is now recognised: they are essential for every citizen to be able to know and critically recognise the context in which they live, and to be able to make decisions, even in situations of uncertainty, on the basis of concrete, shared data.

The national statistical institutes, driven by the conviction that a society in which decisions are taken on the basis of objective data is a society in which democracy reigns, have endeavoured to make access to official data free and easily accessible to their citizens. Free access to statistical information is a means of ensuring the progress of a democratic society, but it loses its effectiveness if citizens are unable to read, interpret and analyse the data. So we were faced with a new challenge: not just disseminating data, but promoting it, getting to know it and making it understood, so that it is useful to the community, a good that has a strong potential impact on everyone's life.

This is why it can be very useful to acquire statistical and data-reading skills at school. Not so much to become statisticians in adulthood, but to learn to read and decode the world around us correctly.

Among the numerous projects aimed at developing statistical literacy in schools, this work aims to illustrate an innovative interdisciplinary teaching pathway designed to promote and develop the principles of active and conscious citizenship in schools, through research activities, the use of open data, the use of information technologies and the civic monitoring of European and national public funding.

The project, aimed at high-schools, develops numerical, statistical and civic education skills, as well as other cross-curricular skills such as critical thinking and problem-solving, teamwork, interpersonal and communication skills, by integrating them into the content of ordinary study subjects, in order to help pupils learn about and communicate, using journalistic techniques, how public policies, and in particular cohesion policies, operate in the places where they live.

Teaching is project-based, combining asynchronous learning moments typical of MOOCs (Massive Online Open Courses) with face-to-face facilitation activities led by the teachers themselves (previously trained ad hoc), group work and online interaction with the project team.

Pupils will therefore discover in class what data is used for, with the help of their teachers, in an interesting and informative journey into the world of figures, which will help them develop their ability to understand social, cultural and environmental developments thanks to the data provided by official statistics. What's more, they will be able to take part in a practical initiative, trying their hand at creating a communication product and putting their analytical skills and creativity to the test.

Keywords. numeracy, statistical literacy, civic mindfulness, active citizenship

1.Introduction

Dans le monde d'aujourd'hui, nous sommes inondés d'informations qui nous parviennent de mille façons (radio, journaux, télévision, internet, etc.), le plus souvent remplies de données statistiques. D6es millions des données transitent sur internet, continuellement mises à jour, retravaillées et présentées avec la meme apparence de crédibilité. Cela pose des problèmes de connaissance et d'orientation, car le citoyen n'est pas toujours en mesure de discerner et de comprendre ce qui se passe dans l'espace public. L'hétérogénéité des modes de collecte, de traitement et de diffusion des données rend leur lecture encore plus difficile.

Dès le début des années 1980, en Italie les rédacteurs des programmes ministériels de l'enseignement de base ont introduit des contenus statistiques dans les programmes scolaires afin de favoriser l'acquisition des compétences nécessaires à la compréhension de l'information quantitative. Aujourd'hui, ces compétences sont indispensables pour naviguer dans les données, savoir évaluer leur qualité, les lire et les utiliser pour prendre des décisions, Il est donc nécessaire d'instaurer une véritable culture statistique à l'école.

Les connaissances statistiques et numériques de base, la capacité à lire, comprendre et évaluer les informations statistiques inhérentes à la vie quotidienne, le raisonnement statistique et la citoyenneté statistique sont les éléments constitutifs de la culture statistique Les élèves sont en effet à la fois les futurs répondants et les futurs utilisateurs des statistiques.. Ces outils peuvent être acquis principalement par le biais de la formation scolaire- Quel meilleur champ d'intervention que les écoles? D'où la décision de considérer les étudiants comme le groupe cible privilégié.

Il s'agit donc d'une alphabétisation visant à l'exercice de compétences critiques et qui ne se limite pas aux matières au sens strict. C'est pourquoi il est nécessaire d'accorder une attention particulière au vocabulaire et à la logique des statistiques. Ce n'est qu'en spécifiant correctement le problème, cette discipline peut aider à trouver des solutions.

Mais comment promouvoir la culture et la compréhension statistiques chez les élèves, sans perdre de vue l'objectif de développer l'esprit critique ?

Le but des instituts statistiques n'est pas seulement de produire , diffuser et analyser des données statistiques, mais aussi d'investir dans l'alphabétisation statistique, autrement dit dans la connaissance des statistiques. La connaissance est en effet l'arme fondamentale pour construire une société ouverte et inclusive, où personne n'est laissé pour compte. La connaissance est l'élément transversal qui sous-tend le changement

2. L'approche de l'enseignement des statistiques

Partant du principe que la réflexion statistique est aussi nécessaire à la citoyenneté active que la capacité à lire et à écrire, ce document relate quelques expériences et initiatives au service de l'enseignement de la statistique comme outil de sensibilisation civique .

La promotion de la culture statistique dans les écoles contribue à l'objectif 4, Cible 4.7 de l'Agenda 2030 : "Faire en sorte que, d'ici à 2030, tous les apprenants acquièrent les connaissances et les compétences nécessaires pour promouvoir le développement durable". connaissances et les compétences nécessaires pour promouvoir le développement durable, notamment par l'éducation au développement et aux modes de vie durables. Le développement durable et les modes de vie durables, les droits de l'homme, l'égalité entre les sexes, la promotion d'une culture pacifique et non violente culture pacifique et non violente, la citoyenneté mondiale et l'appréciation de la diversité culturelle et

de la contribution de la culture au développement durable. de la culture au développement durable" (https://unric.org/it/).

3.**Le project: À l'école pour une citoyenneté active”**

Le project est un parcours didactique interdisciplinaire innovant destiné aux colleges et aux lycées, qui favorise les activités de contrôle civique des financements publics, notamment par l'utilisation de données ouvertes et de technologies de l'information et de la communication.

Le programme d'enseignement du parcours favorise aussi l'acquisition de compétences en matière d'éducation civique, de numérique, de statistiques ,d'éducation civique, de journalisme numérique, et d'autres compétences transversales telles que le développement du sens critique , de la résolution de problèmes, du travail en équipe et des compétences interpersonnelles et de communication, en les intégrant au contenu des matières d'études ordinaires.

L'objectif de chaque équipe d’ètudiants est de mener une recherche thématique pour étudier les caractéristiques socio-économiques, environnementales et/ou culturelles de son territoire à partir d'une intervention financée par les politiques de cohésion sur un thème d'intérêt, choisi sur la base d'informations publiées en format ouvert sur le portail européen vérifiant ainsi comment les politiques publiques interviennent pour améliorer le contexte local.

Le projet est divisé en 4 leçons :

● Planification (leçon 1). Apprendre en quoi consiste le suivi civique, choisir sur le site d' portail de l'Union européenne un projet financé dans votre région à surveiller, identifier une question de recherche. Identifiez une question de recherche, formez le groupe de travail en classe et répartissez-vous en fonction des rôles. Recherchez plus d'informations sur le projet Si le projet est choisi, reconstituez le processus administratif et les décisions publiques qui ont déterminé le projet, identifier les acteurs publics et privés impliqués dans sa mise en œuvre.

● Analyser (leçon 2). Apprendre les techniques de recherche quantitative et qualitative, comprendre ce qu'est l'open data et rechercher des données en rapport avec le sujet choisi, construire un indicateur avec les données trouvées, comprendre le flux de travail. du journalisme de données.

● Explorer (leçon 3). Explorer sur le terrain les progrès du projet choisi par le biais d'une visite. d'un suivi sur place, d'entretiens avec les responsables de la mise en œuvre, de rencontres avec les institutions. Rédiger un rapport de suivi détaillé

● Raconter des histoires (leçon 4). Apprentissage des techniques de communication, conception et mise en œuvre d'une campagne. de sensibilisation et d'engagement pour illustrer les résultats du suivi civique. Organiser un événement public et impliquer la communauté cible pour poursuivre la surveillance sur le projet choisi.

Les classes participant au projet sont soutenues par l'enseignant de référence, avec la collaboration éventuelle d'un enseignant de soutien. Les enseignants et les élèves participent au projet avec le soutien des réseaux territoriaux du project, composés des centres Europe Direct et CDE, des organisations "Amis", des associations à but non lucratif actives dans le domaine des politiques de cohésion, et des personnes de contact territoriales de l'Institut national de la statistique, et tous concourent pour des prix et des récompenses, y compris des voyages éducatifs à Bruxelles auprès des institutions européennes ainsi que des opportunités de formation et d'expérience, en ligne et en présence, et bien plus encore grâce au soutien des nombreux partenaires du projet.Les régions et les

collectivités partenaires apportent leur soutien dans la liaison avec les institutions nationales et locales, avec les autres partenaires du projet (entreprises, fondations, associations) et dans les activités de valorisation du projet à l'échelle nationale**.**

Bibliographie

OECD (2018) Numeracy practices and numeracy skills among adults https:\\www.oecd.org\officialdocuments\publicdisplaydocumentpdf\?cote=EDU\WKP(2018)13&doclanguage=en 5

Carpentieri, J., J. Lister et L. Frumkin (2009), « Adult numeracy: a review of research », http://roar.uel.ac.uk/1340

Hanushek, E. et al. (2015), « Returns to skills around the world: Evidence from PIAAC », European Economic Review, vol. 73, pp. 103-130, http://dx.doi.org/10.1016/j.euroecorev.2014.10.006

Desjardins, R. et K. Rubenson (2011), « An Analysis of Skill Mismatch Using Direct Measures of Skills », Documents de travail de l'OCDE sur l'éducation, no. 63, Éditions OCDE, Paris, http://dx.doi.org/10.1787/5kg3nh9h52g5-en

D.W. Johnson- R.T. Johnson-E.J. Holubec, Apprendimento cooperativo in classe, Erickson,
Catalogue de l'offre Istat-CIC aux écoles– a.s. 2022/2023,

Istat https:\\olimpiadistatistica.istat.it\pluginfile.php\418696\mod_resource\content\5\catalogo.pdf
Dati alla mano https://www.istat.it/it/dati-alla-man

FUTUR PROGRAMME DE MATHÉMATIQUES DU CYCLE 3 : STATISTIQUE ET PROBABILITÉS EN CM1-CM2-SIXIÈME

Charlotte Derouet*[1] and Philippe Dutarte[2]

[1]LISEC UR 2310 – Université de Strasbourg, CNRS – France
[2]Inspection académique – Rectorat de Nice – France

Résumé

Dans cette communication, nous présenterons la partie " *Organisation et gestion de données et probabilités* " du projet de programme du cycle 3 (CM1-CM2-6è), nous pointerons les éléments qui nous semblent problématiques dans ce projet concernant l'enseignement-apprentissage des probabilités et de la statistique, et nous les mettrons en regard avec les éventuels ajustements qui auront été faits, après la consultation nationale, dans la version définitive du programme de cycle 3 (si celle-ci est connue lors des JdS 2025).

*Intervenant

50 NUANCES DE STAT : UNE RÉTROSPECTIVE DES CAFÉS DE LA STATISTIQUE À LYON

Antoine Rolland[1] & Anne-Laure Fougères[2] & Jean-Baptiste Aubin[3]

[1] *Université Lumière Lyon 2, Université Claude Bernard Lyon 1, ERIC, 69007, Lyon, France*
Antoine.rolland@univ-lyon2.fr
[2] *INSA Lyon, CNRS, Centrale Lyon, Universite Claude Bernard Lyon 1, Université Jean Monnet, ICJ UMR5208, 69621 Villeurbanne, France.*
[3] *Universite Claude Bernard Lyon 1, CNRS, Centrale Lyon, INSA Lyon, Université Jean Monnet, ICJ UMR5208, 69622 Villeurbanne, France.*

Résumé. Depuis 2012, une petite équipe propose et organise des « cafés de la statistique » à Lyon. De 4 à 5 séances par an sont proposées (hors période covid) au café La Cloche, situé au centre de Lyon. À l'occasion de la tenue du cinquantième café, nous proposons ici quelques éléments d'histoire et d'analyse de ces rencontres.

Mots-clés. Médiation scientifique, socio-histoire de la quantification

Abstract. Since 2012, a small team has organized "statistics cafés" in Lyon. 4 to 5 sessions per year are offered (excluding covid period) at the La Cloche café, located in the center of Lyon. On the occasion of the fiftieth café, we offer here some elements of history and analysis of these meetings.
Keywords. Scientific mediation, Socio-history of quantification

1 Genèse

Une inspiration : les cafés stat parisiens

Les cafés de la statistique sont nés à Paris en 2005 [1], selon un triple principe :

- Des soirées-débats publiques ouvertes à tous,
- Des débats qui évitent la technicité,
- Des thèmes très variés et d'actualité, un dialogue entre des statisticien·ne·s et le public le plus large : simples citoyen·ne·s, non-spécialistes, représentant.e·s de professions diverses ou participant·e·s au débat politique, économique ou social.

Reprenant ces principes, nous avons lancé les cafés de la statistique à Lyon en mars 2012, autour d'un premier thème, toujours accrocheur en période électorale, sur les sondages politiques. Le choix du lieu s'est porté sur le café La Cloche[1], café historique au centre de Lyon qui à l'époque proposait une dizaine de cafés thématiques (cafés philo, café géo, café de la mode…).

Un objectif : la diffusion de la pensée statistique

L'objectif affiché de l'organisation des cafés de la statistique à Lyon est de participer localement à la diffusion de la pensée statistique. Pour cela, il nous a semblé important de tenir

[1] Deux cafés se sont tenus dans des endroits autres : un dans une librairie, un dans un café proche du lieu des Journées de la Statistique 2022.

ces conférences dans un café de centre-ville, et non dans un local universitaire. Au-delà du fait que cela change la posture de l'intervenant·e, qui se laisse plus facilement interpeller par le public, cela permet surtout à la connaissance de sortir de ses lieux de production (l'université par exemple) pour infuser au plus profond de la société civile. Le soutien de la SFdS, de la Maison des mathématiques et de l'informatique de Lyon (MMI) et de nos laboratoires respectifs est un témoignage du fait que les cafés de la statistique participent de la diffusion et de la médiation scientifique.

Une envie : se faire plaisir en invitant des conférencier.es de qualité

En toute honnêteté, l'organisation des cafés statistique à Lyon est aussi l'occasion pour les organisateurs d'inviter les conférencier·e·s dont ils ont envie d'écouter l'expérience et les connaissances sur tel ou tel sujet. Les sources d'inspirations sont variées : thèmes proposés par le « grand frère » des cafés statistique à Paris, actualité nationale ou locale, tissu local de compétences, suggestions du public…

2 Thèmes & Assistance

Un rythme régulier

Les cafés statistiques lyonnais ont maintenu un rythme régulier de 4 rencontres par an (novembre, janvier, mars, mai), sauf en 2020/2021 pour cause de covid : le café se voulant un lieu de rencontre et d'échanges réels, nous n'avons pas souhaité en organiser en format à distance. Un cinquième café annuel a été organisé par deux fois en marge du colloque Sondage en 2018 et en marge des JdS en 2022.

Des thèmes variés

Les thèmes retenus pour les cafés passés sont très variés, et sont disponibles sur le blog du café [2]. Bien qu'il soit difficile de procéder à des regroupements précis, mentionnons parmi les principaux thèmes abordés lors de ces cafés :

- Les questions de société : mesure de l'illettrisme, de la pauvreté, de la délinquance, du bénévolat, mais aussi les prénoms, les sondages politiques, les processus de vote ou la fouloscopie…
- Les questions de santé : épidémiologie, vaccins, drogue, OGM, génomique…
- Les questions climatiques : changement climatique, biodiversité, transition énergétique…
- La statistique publique : recensement, QPV, indice des prix à la consommation…
- Quelques thèmes plus orientés vers des techniques statistiques précises : Big Data, tests multiples, analyse textuelle, modélisation…
- Enfin bien sûr, plusieurs cafés se sont tenus sur le thème explicite de la sociologie de la quantification ; c'est bien cette approche qui est au cœur de notre démarche : déconstruire la manière dont ont été posées les conventions de mesure dans tel ou tel domaine.

Des intervenant·e·s lyonnais… et d'autres moins

La question des intervenant·e·s disponibles a irrigué depuis le début le programme de notre café : le nombre d'administrations centrales étant par définition moins important à Lyon qu'à Paris, et aucun des initiateurs du café ne provenant du monde de la statistique publique, nous avons d'entrée de jeu couvert un spectre de sujets plus large que les cafés parisiens. La présence

d'une direction régionale de l'INSEE à Lyon nous a permis cependant de pouvoir bénéficier à plusieurs reprises de ces compétences pour des cafés sur la statistique publique.

Nous souhaitions également faire des cafés stats un lieu de rendez-vous de la communauté statistique lyonnaise, complémentaire du réseau des Rencontres Statistiques Lyonnaises (RSL) plus orienté vers les questions de recherches. Ainsi, nous avons sollicité et mis à l'honneur les statisticien·ne·s lyonnais quand les compétences locales étaient disponibles, et nous avons même organisé un certain nombre de séances en partant du choix de l'orateur·rice et non en partant d'un thème spécifique. Le bilan est plutôt positif puisque la moitié des cafés se sont tenus avec des intervenant·e·s locaux, l'autre moitié des intervenant·e·s étant majoritairement parisiens, mais également genevois, toulousain ou grenoblois. Un point à améliorer est le manque de parité : nous n'avons qu'un quart d'intervenantes pour trois quart d'intervenants.

2.1 Un public stable qui se renouvelle

La vocation du café stat est d'être ouvert à tous et toutes : nous ne demandons pas d'inscription au préalable. La diffusion de l'information se fait par l'intermédiaire d'une liste de diffusion de 500 contacts (personnes étant venues une fois et ayant laissé leur adresse mail), mais aussi via le relais de nombreuses autres listes de diffusion (SFdS, MMI, enseignants de math de l'académie, RSL), via LinkedIn également, et nous essayons de diffuser l'information dans les réseaux universitaires possiblement intéressés.

L'effectif moyen d'un café est entre 40 et 50 personnes, plus proche des 40 depuis le covid. Les cafés les plus tardifs (juin) sont généralement un peu moins pleins. Le public est très variable suivant les thèmes ; il est composé d'étudiants ou anciens étudiants en statistique de nos formations respectives, mais aussi souvent de leurs amis et collègues, d'étudiants en sociologie, d'enseignants en mathématique ou retraités, des collègues statisticiens de la région lyonnaise, des habitués du café, mais aussi des personnes venant suite à l'annonce passée dans le journal local comme nous arrivons parfois à le faire. Les personnes reviennent souvent pendant quelques cafés. Une simple règle de trois nous dit que sur un effectif cumulé de 2500 participants sur 50 cafés, la présence de 500 adresses/noms différents montre que les participants viennent à 5 cafés en moyenne, ce qui témoigne d'un bon renouvellement du public.

3 Retour d'expérience

Le retour d'expérience que nous tirons est d'abord celui d'un succès : succès quantitatif, car il n'était pas évident qu'une telle proposition perdure pendant 12 ans ; succès qualitatif également, au vu des intervenant.es que nous avons pu avoir ; succès d'estime enfin, à travers la reconnaissance locale (et nationale !), et une renommée qui dépasse largement le cercle des participants. Bien implanté dans le paysage scientifique les café stat de Lyon sont devenus en 12 ans une institution incontournable du monde statistique et compte bien le rester

Sitographie

[1] https://www.sfds.asso.fr/fr/statistique_et_enjeux_publics/499-les_cafes_de_la_statistique/

[2] https://cafestatistiquelyonnais.blogspot.com/

Annexe : liste des cafés

1 Peut-on croire les sondages préélectoraux? Benoit Riandey (INED) & Bernard Denni (Sciences Po Grenoble)
2 Epidémiologie et insécurité routière Bernard Laumon & Jean-Louis Martin (IFSTTAR)
3 Risques industriels Eric Marsden (Foncsi) & Francois Wahl (IFP)
4 Changement climatique. Aurélien Ribes (Météo France)
5 Recensement : pourquoi, comment Pascal Oger (INSEE)
6 Illettrisme Jean-Pierre Jeantheau (ANLCI)
7 OGM Avner Bar-Hen & Marc Lavielle (Haut Conseil des Biotechnologies)
8 Recensement des SDF Louis Espinasse (INSEE)
9 Marketing Antoine-Eric Sammartino (Lyon 2)
10 Big Data Georges Hebrail (EDF)
11 Secret statistique Jean-Michel Le Gléau (INSEE)
12 Délinquance Laurent Mucchielli (CNRS)
13 Génomique Jean-Baptiste Veyrieras (Biomerieux)
14 Evaluation des politiques publiques Jean-René Brunetière (Pénombres)
15 Statistiques Ethniques Olivier Richomme & Patrick Rozenblatt (Lyon 2)
16 Prévisions à la SNCF Emmanuel Coste (SNCF)
17 Les maths au tribunal Leila Schneps (CNRS)
18 le pôle Rhone alpes de l'orientation Boris François & David Frachisse (PRAO)
19 Bénévolat Cécile Bazin (Recherche et Solidarité)
20 Pauvreté Anaig Abjean (MRIE)
21 Statistiques et enjeux publics Jean Baptiste Aubin et Thibaut Espinasse (INSA et Lyon 1)
22 Drogues Ivana Obradovic (OFDT)
23 Data journalisme Maxime Ferrer (le Monde)
24 Causalité Antoine Chambaz (MAP5)
25 Tests multiples Irène Gannaz (INSA Lyon)
26 Mesure de la biodiversité Joanne Clavel (Paris 8) & Karine Princé (Muséum Histoire Naturelle)
27 Sondages politiques Jean Chiche (Sciences Po)
28 Impact du changement climatique Juliette Blanchet (CNRS Grenoble)
29 Sport Christian Derquennes (SFdS)
30 Prénoms Baptiste Coulmont (Paris 8)
31 ParcoursSup Judicael Courant (lycée La Martinière)
32 IA et Ethique Aurélien Garivier et Philippe Besse (Toulouse)
33 Vaccins Daniel Floret (Haut conseil Santé Publique)
34 Sociologie de la quantification Fabrice Bardet (ENSTP)
35 "il était une fois sur 100", café littéraire Yves Pagès (éditeur)
36 Méthodes de vote Jean-François Laslier (CNRS)
37 Indice des prix à la consommation Jean-Philippe Grouthier (INSEE)
38 Transition énergétique Cédric Philibert (IFRI)
39 Statistique et analyse de texte Adrien Guille (Lyon 2)
40 Epidémiologie Antoine Flahaut (Univ Genève)
41 Quartiers Politique de la ville (QPV) Valérie Darriau (INSEE)
42 Chiffre! Olivier Martin (Paris Cité)
43 Statactivisme Emmanuel Didier (ENS)
44 Complexité Pablo Jensen (ENS Lyon)
45 Mal logement Véronique Gilet (fondation Abbé Pierre)
46 Niveau en math Mélanie Guesnais (Paris Sud)
47 Sagesse ou folie des foules? Aurélien Baillon (EM Lyon)
48 Police scientifique Magalie Faivre & Fanny Guillet (Police Scientifique de Lyon)
49 Surtourisme Victor Anduze (ENSTP)
50 Surprise….

Détection d'anomalies

AGGREGATED MODEL FOR ANOMALY DETECTION: A STATISTICAL LEARNING APPROACH

Rida Kheirallah[1*[0009-0008-8862-2264]] & Anis Hoayek[1] & Mireille Batton-Hubert[1] & Patrick Burlat[2]

[1] *Mines Saint-Etienne, Univ Clermont Auvergne, UMR CNRS 6158 LIMOS, France*
rida.kheirallah@emse.fr
anis.hoayek@emse.fr
batton@emse.fr
[2] *WipSim, Saint-Etienne, France*
patrick.burlat@wipsim.fr

Résumé. Le cœur de cette recherche est un modèle mathématique novateur pour la détection des anomalies dans les processus de fabrication discrète. Conçu pour identifier des écarts significatifs dans les délais de production au sein d'une ligne de production, le modèle est construit en utilisant des techniques d'apprentissage statistique et agrège plusieurs modèles individuels de classification. Il s'appuie sur les principes de la théorie de l'information, notamment l'Entropie de Shannon, pour quantifier l'incertitude et améliorer la précision de l'identification des anomalies. En conséquence, le modèle démontre des performances élevées dans la distinction entre les valeurs aberrantes et les anomalies, ce qui le rend applicable aux données de production réelles. Le modèle est extensible et peut être intégré dans des systèmes de pilotage des lignes de production pour une détection des anomalies en temps réel, s'inscrivant dans les principes de l'Industrie 4.0, notamment la fabrication intelligente et la prise de décision pilotée par les données.

Mots-clés. Détection des Anomalies, Apprentissage Statistique, Théorie de l'Information, Ligne de Production, Système de Pilotage.

Abstract. The core of this research is a novel mathematical model for anomaly detection in discrete manufacturing processes. Designed to identify significant deviations in production lead time within a production line, the model is constructed using statistical learning techniques and aggregates multiple individual classification models. It relies on the principles of information theory, particularly Shannon's Entropy, to quantify uncertainty and enhance the precision of anomaly identification. Consequently, the model demonstrates high performance in distinguishing outliers from anomalies, making it applicable to real-world production data. The model is extensible and can be integrated into production line control systems for real-time anomaly detection, aligning with Industry 4.0 principles, including smart manufacturing and data-driven decision-making.

Keywords. Anomaly Detection, Statistical Learning, Information Theory, Production Line, Control System.

1 Introduction

Anomaly detection, a critical aspect of data analysis, involves identifying patterns or instances that deviate significantly from expected behavior. Widely applied across diverse fields, including healthcare, cybersecurity, and finance, anomaly detection is essential for detecting fraud, system failures, and irregularities that can have significant consequences. According to Chandola et al., (2009), anomaly detection is foundational for understanding and mitigating deviations in complex systems.

In the industrial context, particularly in discrete manufacturing, anomaly detection takes on heightened importance. Modern production lines are highly intricate, driven by interconnected processes and vast streams of real-time data, making the identification of anomalies critical for maintaining

operational efficiency and product quality. As industries transition towards Industry 4.0, the integration of advanced statistical learning models for anomaly detection offers transformative potential by enhancing the precision and speed of identifying significant deviations, thus aligning with the objectives of smart manufacturing and data-driven decision-making. This paper presents a novel approach to anomaly detection tailored to the industrial domain, emphasizing its relevance and applicability in optimizing production line performance.

In the following Section 2, we provide an overview of research, methods, and methodologies for anomaly detection, focusing on production lines and discrete manufacturing environments. Additionally, we conduct in this section a comparative analysis between the aggregated model developed in this work and the models studied in the literature to highlight its advancements and contributions. Section 3 presents the mathematical formalization and construction of the aggregated model, combining Isolation Forest (IF), AutoEncoder (AE), and Support Vector Machine (SVM). Section 4 demonstrates the model's practical application on a real production line database, detailing results through figures and tables. Section 5 summarizes the findings, contributions, and future research directions.

2 State of the Art

Numerous studies have focused on the application of statistical learning methods for anomaly detection and root cause analysis in various industrial contexts. In Zipfel et al., (2023), the authors investigated unsupervised learning models such as Skip-GANomaly, PaDiM, and PatchCore for industrial quality assessment, highlighting their strong performance when compared to supervised methods. They underscored the efficiency of fully unsupervised techniques in identifying anomalies in images. Meanwhile, Chevtchenko et al., (2023) provides a comprehensive review of anomaly detection in industrial machinery using IoT devices and machine learning algorithms. The review examines 84 relevant studies conducted between 2016 and 2023, shedding light on commonly used algorithms, preprocessing methods, and sensor types. It highlights that milling and cutting tools, hydraulic systems, and bearings are the most frequently monitored machinery types, with vibration and temperature sensors being the most prevalent. A variety of machine learning techniques, including supervised, unsupervised, and heuristic approaches, are applied, often combined with preprocessing methods like Fast Fourier Transform and AutoEncoders to enhance accuracy. Nonetheless, challenges remain in performing anomaly detection at the edge, integrating it into existing infrastructure, and adapting machine learning models to evolving industrial environments. Abdelrahman et al., (2020) explores anomaly detection in manufacturing assembly data for two product series, leveraging a range of techniques such as HBOS (Histogram-Based Outlier Score), IF (Isolation Forest), KNN (K-Nearest Neighbors), CBLOF (Clustering-Based Local Outlier Factor), OCSVM (One-Class Support Vector Machine), LOF (Local Outlier Factor), and ABOD (Angle-Based Outlier Detection). Among these, KNN and ABOD demonstrated the best performance. Statistical root cause analysis uncovered seven common rejection causes, with the top three accounting for 85% of rejections. Adjustments made to specifications based on these findings led to a reduction in rejection rates as reported by engineers. Anomaly detection and root cause analysis hold significant importance across various domains beyond supply chains and production lines, including intrusion detection (García-Teodoro et al., 2009) and fraud detection (Pourhabibi et al., 2020) and other domains.

In our review of anomaly detection techniques, we found that most rely on individual supervised or unsupervised models, with no approaches using information theory, such as Shannon Entropy, to handle uncertainty. Our key contribution is a flexible technique that selects models based on criteria like data type, scalability, interpretability, and noise robustness. These selected models are then aggregated to enhance anomaly detection performance, addressing diverse challenges across contexts. Our approach integrates IF, SVM, and AE models, chosen for their popularity and suitability for numerical data. Table 1 compares these models with the aggregated approach, using performance indicators where "0" means the criterion is unmet, "1" means it is met, and "(1)" indicates non-binary classification. The aggregated model excels in outlier detection due to its use of Shannon Entropy and adaptability, outperforming individual models. Developed in Python, the model integrates easily into industrial control systems, leveraging Python's compatibility and built-in optimizations for practical use in various industrial

settings.

Criteria	IF	SVM	AE	Agg Model
Capacity Distinguishing Anomalies/Outliers	0	0	0	1
Scalability	1	0	(1)	1
Interpretability	1	(1)	0	1
Handles Diverse Data Types	(1)	(1)	1	1
Noise Robustness	1	0	(1)	1

Table 1: Performance comparison between individual models and the aggregated model

3 Aggregated Model for Anomaly Detection

This section outlines the mathematical formulation of the aggregated anomaly detection model, which combines IF, SVM, and AE classifiers. Each model is trained on the dataset (Train_Data), and standardized scores are calculated for each observation. Entropy is then calculated from the scores to measure uncertainty. Predefined anomaly detection thresholds are established based on quantiles of the standardized scores and standardized Entropy. The model detects anomalies in Test_Data by flagging observations with scores and Entropy that surpass the predefined thresholds, providing a robust detection method with customizable thresholds for various data types.

3.1 Mathematical Formulation

The IF algorithm detects anomalies by isolating instances in a dataset through random partitioning in a binary tree. Shorter average path lengths in the tree indicate anomalies. The isolation score, which measures the average path length, identifies instances with lower scores as more likely to be anomalies (Liu et al., 2012).

SVMs used as one-class classifiers for anomaly detection, find a hyperplane that maximizes the margin between normal instances and potential anomalies. The SVM optimizes a cost function to separate these instances, and the distance from the hyperplane serves as an anomaly score (Cortes et al., 1995).

AEs are neural networks that learn a compact representation of input data by encoding and reconstructing it. The reconstruction error, typically measured by Mean Squared Error, is used as an indicator of anomalies (Tschannen, et al., 2018).

3.2 Model Construction

3.2.1 Model Training and Anomaly Score

Each anomaly detection model is separately trained on the train dataset (Train_Data), and assigns an anomaly score to each observation. These scores are then standardized to the interval $[0,1]$, with scores closer to one indicating a higher likelihood of abnormality. Standardizing scores allows for consistent and fair comparisons across models, enhancing interpretability and supporting robust evaluation of each model's performance.

3.2.2 Anomaly Score into Entropy

For each observation in Train_Data, the three standardized scores calculated are first transformed into a probability distribution. Thus, for a given observation of Train_Data, after calculating the following three standardized scores: $score_{SVM}, score_{AE}$ and $score_{IF}$, the probability corresponding to the SVM model (denoted as prb_{SVM}) is calculated using the following Equation :

$$prb_{SVM} = \frac{score_{SVM}}{score_{SVM} + score_{AE} + score_{IF}} .$$

Similarly, prb_{AE} and prb_{IF} are calculated in the same manner. Then after calculating prb_{SVM}, prb_{AE} and prb_{IF}, we calculate from these probabilities the *Entropy* of the observation using Equation 1:

$$Entropy = -\sum_{j=1}^{j=3} prb_j \times log_2(prb_j), \quad (1)$$

where $prb_1 = prb_{SVM}$, $prb_2 = prb_{AE}$ and $prb_3 = prb_{IF}$.

After the calculation of the Entropy, it is normalized and corporated into the anomaly detection model criteria to strengthen robustness and precision. High entropy values (near 1) indicate alignment among the three models (IF, SVM, AE), where each assigns nearly equal probabilities (around 1/3), confirming that the three models agree about the nature of the analyzed observation (anomalous or normal). In contrast, low entropy values (close to 0) signal disagreement, e.g. one model strongly detecting an anomaly while the others do not. This entropy-driven approach enhances the system's capacity to distinguish true anomalies (i.e. when all the models are aligned about considering the abnormality nature of the observation) from outliers (i.e. when a strong disagreement exists among the different models about the abnormality nature of the observation) by identifying intricate patterns beyond extreme values, thereby improving detection accuracy and reliability.

3.2.3 Identification of Quantiles

In this step, based on Tain_Data, four predefined threshold percentages are set: IF, SVM, AE, and Entropy thresholds. These thresholds are used to determine the quantiles for each standardized score and the standardized Entropy. Once calculated, these quantiles are recorded. Additionally, the three trained anomaly detection models are saved for future use in identifying anomalies within a test dataset called Test_Data.

3.2.4 Anomaly Detection

For each observation in Test_Data, the three standardized scores and the standardized Entropy are computed using the trained models. An observation is classified as an anomaly if it meets both of the following criteria:

1. The standardized scores exceed their respective predefined quantile values.
2. The standardized Entropy exceed its predefined quantile value.

4 Industrial Application in a Production Line

4.1 Context of Anomalies

Before presenting the results and detected anomalies, it's essential to establish the context of our work. Figure 1 illustrates the estimated production lead times (PLTs) of a production order (PO) at a workstation on the production line. The PLT represents the total time a PO spends at the workstation during the fabrication process. This includes both the fabrication time (the actual time spent manufacturing the PO) and the wait time (the time spent in queues at the workstation). These PLTs are estimated before the completion of the PO's manufacturing process on that workstation. Each PLT estimation is associated with one of 20 archives, enabling tracking of these estimations over time. In Figure 1, the X-axis represents archive numbers (earliest to latest), while the Y-axis shows PLT estimates in hours. Initially, the estimated PLT is 3.1 hours, increasing to 12.5 hours by the final archive, converging toward the actual PLT of 11.6 hours observed after processing. This convergence reflects a Markovian process, where each new estimation incorporates prior estimates, enabling continuous refinement and improved accuracy.

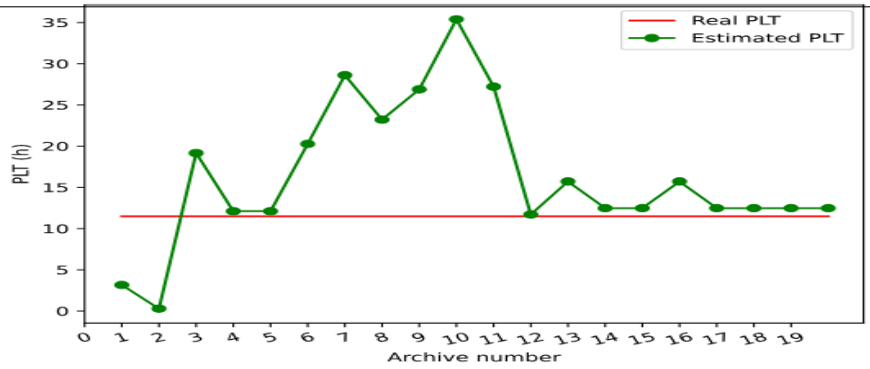


Figure 1 : Real PLT VS Estimated PLT of a PO on a WorkStation

The estimated PLTs of a PO on a workstation WS_i (where $i \in \{1, \dots, L\}$ and L is the total number of workstations in the production line **,** can be represented as a vector denoted $PLT^{E}_{(PO,WS_i)} \in \mathbb{R}^K$, such that,

$$PLT^{E}_{(PO,WS_i)} = (PLT^{E}_{(1)_{(PO,WS_i)}}, \dots , PLT^{E}_{(j)_{(PO,WS_i)}}, \dots , PLT^{E}_{(K)_{(PO,WS_i)}}),$$

Where K is the total number of archives, $PLT^{E}_{(1)_{(PO,WS_i)}}$ is the estimated PLT of the PO at WS_i recorded from the first archive at time $(t_1)_{(PO,WS_i)}$, $PLT^{E}_{(K)_{(PO,WS_i)}}$ is the PLT of the PO at WS_i, recorded from the last archive at time $(t_K)_{(PO,WS_i)}$ and $PLT^{E}_{(j)_{(PO,WS_i)}}$ is an intermediate estimated PLT of the PO at WS_i, recorded from an intermediate archive at time $(t_j)_{(PO,WS_i)}$.
Denoting $PLT^{R}_{(PO,WS_i)}$ as the actual PLT of the PO on WS_i, we then define the Error vector $E_{(PO,WS_i)} \in \mathbb{R}^K$, such that,

$$E_{(PO,WS_i)} = (E_{(1)_{(PO,WS_i)}}, \dots, E_{(j)_{(PO,WS_i)}}, \dots , E_{(K)_{(PO,WS_i)}}),$$

where $E_{(j)_{(PO,WS_i)}}$ is the absolute value of the difference between $PLT^{R}_{(PO,WS_i)}$ and $PLT^{E}_{(j)_{(PO,WS_i)}}$.
Thus, $E_{(j)_{(PO,WS_i)}}$ represents the error of the estimated PLT corresponding to the archive recorded at time $(t_j)_{(PO,WS_i)}$. After defining $E_{(PO,WS_i)}$, we consequently calculate from this vector its statistical median denoted as $Median\ (E_{(PO,WS_i)})$ and given by:

$$Median\left(E_{(PO,WS_i)}\right) = PLT^{R}_{(PO,WS_i)} - Median\left(PLT^{E}_{(PO,WS_i)}\right),$$

where $Median\left(PLT^{E}_{(PO,WS_i)}\right)$ is the statistical median of $PLT^{E}_{(PO,WS_i)}$.
We then define the *Input Error Vector* (IEV) for a PO at WS_i denoted as: $IEV_{(PO,WS_i)} \in \mathbb{R}^3$, such that,

$$IEV_{(PO,WS_i)} = \left(E_{(1)_{(PO,WS_i)}}, Median\left(E_{(PO,WS_i)}\right), E_{(K)_{(PO,WS_i)}} \right). \quad (2)$$

$IEV_{(PO,WS_i)}$ will be injected into the different anomaly detection models. The concept of anomaly of a PO at a WS_i is defined through the statistical analysis of the $IEV_{(PO,WS_i)}$. A PO is considered to exhibit abnormal behavior at WS_i if $f\left(IEV_{(PO,WS_i)}\right) = 1$, where,

$$f : \mathbb{R}^3 \rightarrow \{0, 1\},$$
$$IEV_{(PO,WS_i)} \rightarrow f\left(IEV_{(PO,WS_i)}\right) = \begin{cases} 0 \Rightarrow PO \text{ is normal at } WS_i \\ 1 \Rightarrow PO \text{ is abnormal at } WS_i \end{cases} \quad (3)$$

The function f defined by Equation 3 is the heart of our work, representing the model we built by aggregating multiple individual classification model and incorporating Shannon's Entropy.

4.2 About Database

The analyzed database from WipSim (https.//wipsim.fr), a company specializing in production

line control systems, contains 4 GB of data with 516 archives recorded between February 6 and July 27, 2024. It tracks 1008 POs across 21 workstations for boring machine components, providing actual and estimated PLTs, with an average of three archives per day.

4.3 Model Application

The model's performance is demonstrated using Train_Data and Test_Data, which correspond to the workstation Conwip2_12_6 of the production line. AppendixBTable1 lists 65 POs from Train_Data, detailing their IEV values (Equation 2), individual model scores, normalized Entropy (Equation 1), and 90% anomaly quantile thresholds. After training, anomalies are detected in Test_Data. In AppendixAFigure1, we present a 3D visualization of the Test_Data observations, highlighting the abnormal observation in red and normal observations in blue. The IEV values of these observations are listed in AppendixBTable2. Key observations include PO with ID equal 258418, which was flagged as anomalous due to significantly higher error metrics, and PO with ID equal 258372, which highlights the benefits of aggregation. Without aggregation, this observation would have been flagged as abnormal in a single IF model due to its IF score equal to one. However, the aggregated model identifies it as normal because its Entropy is 0.88, which is below the 90% threshold of 0.92 (AppendixBTable1).
To demonstrate the model's capability to differentiate outliers from anomalies, Test_Data from workstation Conwip2_8_6 is analyzed. In AppendixAFigure2, we highlight in red the abnormal PO detected by the model and in green we mark the outlier. As depicted in this figure, despite the proximity of these two observations, only one was identified as abnormal. AppendixBTable3 provides the IEVs of Test_Data for Conwip2_8_6.

5 Conclusion

This paper explores anomaly detection in industrial production lines, proposing a novel model that aggregates multiple individual classification models with Shannon's Entropy. Applied to a real production line, the model effectively detects abnormal POs at workstations, highlighting strong performance in distinguishing anomalies from outliers, scalability, and seamless dashboard integration. This approach supports proactive anomaly management and aligns with Industry 4.0's vision of smart manufacturing. Future work involves expert labeling of POs to optimize model parameters and refine anomaly thresholds.

References

Chandola, V., Banerjee, A., & Kumar, V. (2009), Anomaly detection: A survey, *ACM Computing Surveys*, 41(3), 1-58.

*Zipfel, J., Verworner, F., Fischer, M., Wieland, U., Kraus, M., & Zschech, P. (*2023). Anomaly detection for industrial quality assurance: A comparative evaluation of unsupervised deep learning models, *Computers & Industrial Engineering*.

Chevtchenko, S. F., Rocha, E. d. S., Dos Santos, M. C. M., Mota, R. L., Vieira, D. M., de Andrade, E. C., & Araújo, D. R. B. (2023**).** Anomaly Detection in Industrial Machinery using IoT Devices and Machine Learning, *IEEE Access, 11, 128288-128305*.

Abdelrahman, A., & Keikhosrokiani, P. (2020). Assembly Line Anomaly Detection and Root Cause Analysis Using Machine Learning, IEEE Access, 8, 189661-189672. doi: 10.1109/ACCESS.2020.3029826

*García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (*2009). Anomaly-based network intrusion detection: Techniques, systems, and challenges, *Computers & Security*, 28(1–2), 18–28.

Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches, *Decision Support Systems*, 133, 113303.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection, ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1), 1-39

Cortes, C., & Vapnik, V. (1995). Support-vector networks, Machine Learning.

Tschannen, M., Bachem, O., & Lucic, M. (2018). Recent Advances in Autoencoder-Based Representation Learning. *arXiv preprint arXiv:1812.05069*.

Density Based Isolation Forest

Nathan Leveque[1,2] & Emmanuel Frénod[1] & Jérôme Lacaille[2]
Audrey Poterie[1] & François Septier[1] & Mourad Yahia Bacha[2]

[1] *Université Bretagne Sud, UMR CNRS 6205, LMBA, F-56000 Vannes, France*
[2] *Safran Aircraft Engine, France*

Resumé. Ce travail porte sur la détection d'anomalies. Nous proposons une nouvelle méthode de forêt d'isolation que nous appelons la forêt d'isolation basée sur la densité (dbiForest). Comme les forêts d'isolation (iRF) introduites par Liu *et al.* (2008), une forêt dbiForest se compose de plusieurs arbres d'isolation basés sur la densité (notés dbiTree). Chaque dbiTree cherche à identifier les anomalies au moyen de divisions binaires et récursives de l'espace des données. L'originalité de cette nouvelle approche repose sur l'utilisation d'estimations de densités pour estimer les coupures. L'algorithme s'appuie sur l'hypothèse que, les anomalies correspondant à des observations à la fois différentes et rares par rapport aux données normales, ces observations seraient présentes dans des espaces de plus faibles densités. Chaque dbiTree cherche donc à isoler les observations présentes dans les régions de plus faible densité. Nous montrons les bonnes performances de notre méthode au travers de plusieurs expériences numériques dans lesquelles dbiForest est comparé à d'autres méthodes, notamment l'algorithme iRF.

Mots-clés. Détection d'anomalies, Forêt d'Isolation, Estimation de densité par noyau.

Abstract. This work focuses on anomaly detection. We propose a new isolation forest method that we call the density-based isolation forest (dbiForest). Like the isolation forests (iRF) introduced by Liu *et al.* (2008), a dbiForest consists of several density-based isolation trees (denoted dbiTree). Each dbiTree seeks to identify anomalies by means of binary and recursive divisions of the data space. The originality of this new approach lies in the use of density estimations to estimate the splits. The algorithm is based on the assumption that, since anomalies correspond to observations that are both different and rare compared to normal data, these observations would be present in spaces of lower densities. Each dbiTree therefore aims to isolate observations that lie in regions of lower density. We show the good performances of our method through several numerical experiments in which dbiForest is compared with other methods, notably the iRF algorithm.

Keywords. Outlier detection, Isolation Forest, Kernel density estimation.

1 Introduction

On se place dans le cadre de l'apprentissage non supervisé où les données consistent en un échantillon $\mathbb{X} = \{\mathbf{X}_i\}_{i=1}^n$ composé de n copies du vecteur aléatoire $\mathbf{X} = (X^1, \cdots, X^p)$ défini sur $\mathcal{X}$. Pour simplifier, on considère $\mathcal{X} = \mathbb{R}^p$. Dans ce contexte, on s'intéresse à la détection d'anomalies. D'après Ghalia *et al.* (2023), une anomalie, en anglais *outliers*, peut

être définie comme une observation qui se produit très rarement dans un échantillon et dont les caractéristiques diffèrent significativement de celles observées pour la plupart des données de l'échantillon. La détection de ces points anormaux est un sujet largement étudié et qui reste encore aujourd'hui un enjeu majeur dans beaucoup de domaines, notamment en industrie. De nombreuses méthodes statistiques ont été développées pour répondre à ce problème. Un état de l'art est par exemple présenté par Chabchoub *et al.* (2022). Dans ce travail, nous nous sommes plus précisément intéressés à l'algorithme des forêts aléatoires d'isolation (en anglais *isolation forest*, iRF) introduit par Liu *et al.* (2008). Cette méthode non supervisée fait partie des approches en apprentissage statistique couramment utilisées pour la détection d'anomalies. Une forêt iRF se compose d'un grand nombre $K \in \mathbb{N}^*$ d'arbres d'isolation (iTree). L'algorithme commence par construire K d'échantillons bootstrap indépendants, notés $\mathbb{X}^1, \cdots, \mathbb{X}^K$. Chaque échantillon $\mathbb{X}^k$ est obtenu en tirant aléatoirement et sans remise $n' < n$ observations de $\mathbb{X}$. Chaque échantillon $\mathbb{X}^k$ $(k = 1, \cdots, K)$ est ensuite utilisé pour construire un iTree, noté T^k $(k = 1, \cdots, K)$. Les iTree cherchent à isoler les points anormaux au moyen de divisions binaires et itératives de l'espace $\mathcal{X}$. Une coupure est définie par le choix uniforme d'une direction d et d'un seuil s_d pour celle-ci. La première étape de construction de l'iTree T^k consiste à diviser $\mathcal{X}$ en deux sous-espaces disjoints :

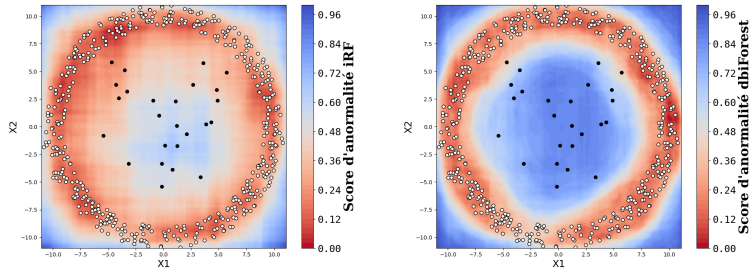
$$\{\mathbf{X}_i \in \mathbb{X}^k \mid X_i^d < s_d\} \quad \text{et} \quad \{\mathbf{X}_i \in \mathbb{X}^k \mid X_i^d \geq s_d\}. \tag{1}$$

Les deux sous-espaces ainsi obtenus sont appelés des nœuds. La procédure de découpage est ensuite répétée sur chacun de ces deux nœuds, puis de manière récursive sur tous les nœuds résultants jusqu'à obtenir des nœuds unitaires ou atteindre un nombre de découpe maximal fixé. À la fin, on obtient donc une partition de l'espace $\mathcal{X}$ qui peut être représentée sous la forme d'un arbre, appelé arbre iTree maximal. À chaque iTree T^k (avec $k = 1, \cdots, K$) et chaque observation $\mathbf{X}_i \in \mathbb{X}$ (avec $i \in [\![1, n]\!]$), on associe un entier positif $p_k(\mathbf{X}_i)$ défini comme la profondeur du nœud terminal de T^k contenant $\mathbf{X}_i$. Ainsi, à partir des K arbres de la forêt d'isolation, on peut estimer le score d'anormalité de chaque observation $\mathbf{X}_i \in \mathbb{X}$:

$$\mathcal{S}(\mathbf{X}_i) = 2^{\frac{-1}{c(n)} \frac{\sum_{k=1}^{K} p_k(\mathbf{X_i})}{K}}, \tag{2}$$

où $c(n)$ correspond à une estimation de la profondeur moyenne pour atteindre un nœud terminal (expression analytique donnée par Bruno (1999)). L'idée de l'algorithme iRF repose sur l'hypothèse que les anomalies étant plus rares et très différentes des observations normales, elles sont isolées plus rapidement dans les arbres iTree. Ainsi dans un forêt iRF, les anomalies correspondent aux observations ayant les scores d'anormalité les plus élevés. L'algorithme détermine un seuil d'anormalité $\mathcal{S}_a$ pour $\mathcal{S}$ à partir de la proportion d'anomalies ϵ, supposée connue. Ce seuil $\mathcal{S}_a$, qui est défini comme le quantile de niveau $1 - \epsilon$ de la série des scores estimés $\{\mathcal{S}(\mathbf{X}_1), \cdots, \mathcal{S}(\mathbf{X}_n)\}$, est utilisé pour identifier les anomalies dans l'échantillon $\mathbb{X}$. Il permet aussi de savoir si une nouvelle observation doit être considérée comme une anomalie ou non. En effet, une observation $\mathbf{X} \in \mathcal{X}$ sera considérée comme une anomalie si son score d'anormalité vérifie la condition : $\mathcal{S}(\mathbf{X}) \geq \mathcal{S}_a$. L'algorithme iRF possède de nombreux atouts comme des hypothèses de validité très peu restrictives sur la distribution des données, une implémentation facile et un faible coût computationnel. De plus, la méthode a montré de bonnes performances dans de nombreux domaines (Al-Shehari *et al.* 2023, Eugen Stripling *et al.* 2018). Néanmoins, ses performances dépendent notamment de la géométrie des données

(Sahand *et al.*, 2021). Considérons l'exemple simple de la figure 1a. Dans ce dernier, les points normaux (en blanc) sont distribués uniformément sur une couronne de rayon intérieur 9 et de rayon extérieur 11, tandis que les points anomaux (en noir) sont distribués uniformément sur un carré centré en (0,0) et de côté de longueur 6. La Figure 1a montre la carte des scores d'anormalité estimés avec une forêt iRF. Sur cet exemple, on peut remarquer que la méthode échoue en affectant des scores d'anormalité similaires aux observations anormales et normales (F1-score= 44%). Ce résultat peut s'expliquer notamment par l'utilisation de coupures linéaires et unidirectionnelles non adaptées dans l'exemple ci-dessous. Dans ce contexte, différents algorithmes de forêts d'isolations ont été développés. On peut citer, par exemple, les forêts d'isolation étendues (eiForest) introduites par Sahand *et al.* (2021) qui découpent l'espace des données en utilisant des hyperplans de dimensions variables. Cette dernière permet dans certains cas d'obtenir de meilleurs résultats que l'iRF. Néanmoins, comme l'iRF, ses performances restent très dépendantes de la géométrie des données. Dans ce travail, nous proposons un nouvel algorithme de forêts d'isolation qui utilise des estimations de densités pour partitionner l'espace et donc isoler les anomalies. Cette nouvelle approche que nous appelons dbiForest (pour forêt d'isolation basée sur la densité) est présentée dans la section suivante.



(a) iRF (F1-score= 44%) (b) iForest (F1-score= 88%)

FIGURE 1 – Carte du score d'anormalité de la forêt iRF (a) et de la forêt dbiForest (b). *Note de lecture : Les points normaux sont en blanc et les anormaux sont en noir.*

2 La dbiForest

L'algorithme dbiForest peut être vu comme une extension des forêts d'isolation iRF. Une forêt dbiForest est composée d'un grand nombre K d'arbres d'isolation basés sur la densité et notés dbiTree. L'algorithme génère indépendamment K échantillons bootstrap $\mathbb{X}^1, \cdots, \mathbb{X}^K$ en tirant aléatoirement et sans remise $n' < n$ observations dans $\mathbb{X}$. Un arbre dbiTree est ensuite construit sur chaque échantillon bootstrap. Comme pour les forêts d'isolation, chaque dbiTree cherche à isoler les points anormaux au moyen de divisions binaires et itératives de l'espace des données $\mathcal{X}$. Néanmoins, contrairement aux forêts d'isolation iRF, dans un arbre dbiTree, les coupures sont définies à partir d'estimations par noyaux de la densité des données. De plus, les coupures ne sont pas choisies aléatoirement. Considérons la première étape de construction de l'arbre T^k (avec $k = 1, \cdots, K$) qui consiste à diviser l'espace tout entier des variables $\mathcal{X}$ en deux nœuds disjoints. Pour ce faire, l'algorithme tire uniformément $n_{try} \in$

$\mathbb{N}$ combinaisons de variables, notées $D_1, D_2, ..., D_{n_{try}}$, dans l'ensemble $\{1, \cdots, p\}$. La taille de chaque combinaison est choisie aléatoirement dans l'intervalle $[D_{\min}, D_{\max}]$. Les hyper-paramètres $D_{\min}$ et $D_{\max}$ sont deux entiers positifs fixés en amont et avec $D_{\min} < D_{\max} << p$. Ensuite, pour chaque combinaison de variables D_h $(h = 1, \cdots, n_{try})$, l'algorithme estime la densité $\hat{f}^{k,D_h}$ sur l'échantillon $\mathbb{X}^k$ à l'aide de noyaux gaussiens :

$$\hat{f}^{k,D_h}(\mathbf{X}) = A \sum_{j=1}^{n'} e^{-\frac{1}{2}(\mathbf{X}_j^{k,D_h} - \mathbf{X}^{D_h})^T H^{-1} (\mathbf{X}_j^{k,D_h} - \mathbf{X}^{D_h})}, \tag{3}$$

où $\mathbf{X} \in \mathcal{X}$, $A = \frac{1}{n'}(2\pi)^{\frac{-Card(D)}{2}}|H|^{-\frac{1}{2}}$, $H \in \mathcal{M}_{D_h \times D_h}(\mathbb{R}^+)$ représente la fenêtre du noyau sélectionnée en utilisant la règle du pouce de Scott et $\mathbb{X}^{k,D_h} = \{\mathbf{X}_i^{D_h}\}_{i=1}^{n'}$ désigne la restriction de $\mathbb{X}^k$ aux variables de l'ensemble D_h. Le seuil s_{D_h} utilisé pour la coupure est tiré uniformément dans l'intervalle $\left[\min_{i \leq n'}(\hat{f}^{k,D_h}(\mathbf{X}_i)), \max_{i \leq n'}(\hat{f}^{k,D_h}(\mathbf{X}_i))\right]$. Pour découper $\mathcal{X}$, l'algorithme sélectionne un couple (D^*, s_{D^*}) parmi les n_{try} candidats possibles. Par définition, les anomalies sont supposées être des observations rares et différentes des données normales. Autrement dit, on peut supposer que les anomalies sont des observations présentes dans les régions de $\mathcal{X}$ de faibles densités tandis que les points normaux sont davantage présents dans des régions denses de $\mathcal{X}$. Utilisant cette hypothèse, l'algorithme choisit la combinaison de variables qui reflète le mieux le passage d'un espace de très faible densité à un espace de fortes densités. Pour ce faire, pour chaque combinaison D_h $(h = 1, \cdots, n_{try})$, l'algorithme estime la densité $\hat{f}^{k,D_h}(\mathbf{X}_i)$ pour toute observation $\mathbf{X}_i \in \mathbb{X}^k$. Ces valeurs sont ensuite ordonnées dans l'ordre croissant pour définir les accroissements point à point :

$$\Delta_{(l)}\hat{f}^{k,D_h} = \hat{f}_{(l+1)}^{k,D_h} - \hat{f}_{(l)}^{k,D_h}, \tag{4}$$

où $\hat{f}_{(l)}^{k,D_h}$ est la l-ème plus petite valeur dans la séquence $\{\hat{f}^{k,D_h}(\mathbf{X}_1), \cdots, \hat{f}^{k,D_h}(\mathbf{X}_{n'})\}$, $\Delta_{(l)}\hat{f}^{k,D_h}$ désigne le l-ème accroissement pour $\hat{f}^{k,D_h}$ et avec $l = 1, \cdots, n' - 1$. Chaque accroissement est normalisé en divisant par la somme totale des accroissements :

$$\Delta_{(l)}^n \hat{f}^{k,D_h} = \frac{\Delta_{(l)}\hat{f}^{k,D_h}}{\sum_{l'=1}^{n'-1} \Delta_{(l')}\hat{f}^{k,D_h}}. \tag{5}$$

L'algorithme sélectionne finalement la coupure (D^*, s_{D^*}) associée au plus fort accroissement :

$$D^* = \underset{D_h \in \{D_1, \cdots, D_{n_{try}}\}}{\operatorname{argmax}} \left(\max_{l \in \{1, \cdots, n'-1\}} \left(\Delta_{(l)}^n \hat{f}^{k,D_h} \right) \right). \tag{6}$$

La première étape de construction de T^k consiste à diviser l'espace $\mathcal{X}$ en les deux nœuds :

$$\{\mathbf{X}_i \in \mathbb{X}^k \mid \hat{f}^{k,D^*}(\mathbf{X}_i) < s_{D^*}\} \quad \text{et} \quad \{\mathbf{X}_i \in \mathbb{X}^k \mid \hat{f}^{k,D^*}(\mathbf{X}_i) \geq s_{D^*}\}. \tag{7}$$

La procédure de découpage est ensuite répétée de manière itérative sur les nœuds ainsi créés, et ce, jusqu'à ce que tous les nœuds terminaux contiennent moins de $D_{\max} + 1$ observations (condition requise pour estimer les densités). Une fois les K arbres dbiTree construits, l'algorithme dbiForest associe, à chaque point de l'échantillon $\mathbb{X}$, la profondeur moyenne, notée

$\bar{p}(\mathbf{X}_i)$, qui est définie comme la moyenne, sur les K arbres de la forêt, des profondeurs des nœuds terminaux contenant l'observation. À partir de cette profondeur moyenne, la méthode peut calculer le score d'anormalité (2). dbiForest détermine également un seuil d'anormalité $\bar{p}_a$ pour la profondeur moyenne et en utilisant, contrairement aux forêts iRF, les échantillons out-of-bag. Pour ce faire, l'algorithme estime, pour chaque observation $\mathbf{X}_i \in \mathbb{X}$, sa profondeur moyenne sur les échantillons out-of-bag comme suit :

$$\bar{p}_{OOB}(\mathbf{X}_i) = \frac{\sum_{k=1}^{K} p_k(\mathbf{X}_i)\mathbb{1}(\mathbf{X}_i \in \mathbb{X}_{OOB_k})}{\sum_{k=1}^{K} \mathbb{1}(\mathbf{X}_i \in \mathbb{X}_{OOB_k})}, \tag{8}$$

où $OOB_k = \mathbb{X} \setminus \mathbb{X}^k$ désigne le k-ème échantillon out-of-bag associé et $\mathbb{1}(\mathcal{C})$ correspond à la fonction indicatrice qui vaut 1 si la condition $\mathcal{C}$ est respectée et 0 sinon. On peut noter que $\bar{p}_{OOB}(\mathbf{X}_i)$ existe si et seulement si $\mathbf{X}_i \in \mathbb{X}$ est dans au moins un échantillon out-of-bag. On détermine ensuite le seuil $\bar{p}_a$ défini comme le quantile de niveau ϵ de la séquence des profondeurs moyenne $\bar{p}_{OOB}$ estimées sur $\mathbb{X}$. C'est ce seuil d'anormalité qui est ensuite utilisé pour identifier les anomalies de notre échantillon et déterminer si une nouvelle observation est une anomalie ou non. En effet, une nouvelle observation $\mathbf{X} \in \mathcal{X}$ sera considérée comme une anomalie si elle vérifie la condition : $\bar{p}(\mathbf{X}) \leq \bar{p}_a$. La section suivante présente quelques résultats numériques basés sur le modèle présenté Figure 1.

3 Résultats numériques

La Figure 1(b) présente les résultats obtenus avec l'algorithme dbiForest dans l'exemple présenté précédemment. On peut remarquer que contrairement à l'algorithme iRF, notre méthode attribue des scores d'anormalité bien plus élevés aux points anormaux (points noirs) qu'aux points normaux (points blancs) amenant à un F1-score bien plus important de 88%. Nous avons ensuite généré 500 échantillons de données selon le même modèle en faisant varier aléatoirement et uniformément la longueur du côté du carré (entre 1 et 3) ainsi que les rayons intérieurs (entre 5 et 15) et extérieurs (entre 110 % et 120 % de la valeur du rayon intérieur) de la couronne. L'utilisation de différentes valeurs pour les rayons et le côté du carré permettent de considérer des scénarios de complexité variable. L'algorithme iRF et notre méthode dbiForest ont ensuite été appliqués sur chacun des jeux de données. Pour chaque itération et chaque méthode, nous avons mesuré le F1-score qui est un indicateur permettant de faire un compromis entre la capacité à détecter les anomalies et à éviter les faux positifs. Les résultats sont visibles sur la Figure 2. On peut remarquer que, dans cet exemple, notre méthode surpasse largement la méthode iRF avec des résultats à la fois meilleurs et plus stables.

4 Conclusion

Nous avons présenté ici une nouvelle méthode originale et prometteuse de forêts d'isolation, la dbiForest, qui permet, grâce à l'utilisation d'estimations de densités, de détecter

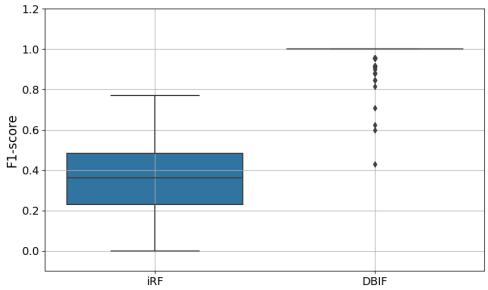


FIGURE 2 – Comparaison de l'algorithme de l'iRF et des dbiForest

des anomalies dans des cas où la forêt d'isolation échoue. Ce résumé présente ce nouvel algorithme de façon assez succincte et montre quelques illustrations. Néanmoins, des analyses de simulations approfondies ainsi que des applications sur des données réelles issues de l'industrie aéronautique ont aussi été réalisées. D'autre part, le lecteur aura peut-être noté quelques similitudes entre notre approche et l'algorithme DBSCAN (Ester *et al.*, 1996), une des méthodes statistiques de référence pour la détection d'anomalies qui s'appuie sur des estimations locales de la densité. Nos analyses intègrent aussi une comparaison avec cette méthode. Les résultats pourront être présentés lors de la conférence.

5 Bibliographie

Liu, Fei Tony Ting, Kai Ming and Zhou, Zhi-Hua (2008), *Isolation Forest* Eighth IEEE International Conference on Data Mining

Sahand Hariri, Matias Carrasco Kind, Robert J.Brunner (2021), *Extended Isolation Forest*, IEEE Transactions on Knowledge and Data Engineering, pp.1479-1489

Chabchoub Y, Togbe M, Boly A, Chiky R, (2022), *An In-Depth Study and Improvement of Isolation Forest*, IEEE Access, 5, pp. 10219-10237

Falahi, Thaer, Nasserddine, Ghalia, Younis, Joumana (2023) *Detecting Data Outliers with Machine Learning*, Al-Salam Journal for Engineering and Technology

T. Al-Shehari, M. Al-Razgan, T. Alfakih, R. A. Alsowail and S. Pandiaraj (2023) *Insider Threat Detection Model Using Anomaly-Based Isolation Forest Algorithm* IEEE Access, vol. 11, pp. 118170-118185

Eugen Stripling, Bart Baesens, Barak Chizi et Seppe vanden Broucke (2018) *Isolation-based conditional anomaly detection on mixed-attribute data to uncover workers' compensation fraud* Decision Support Systems 111 pp. 13-26

Ester M, Kriegel H, Sander J, Xu X (1996) *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise* Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining pp. 226-231

Bruno R. Preiss (1999) *Data Structures and Algorithms*

Tester la dynamique maligne des biomarqueurs tumoraux à l'aide des signatures d'un chemin rugueux.

Rémi Vaucher [1,2] & Stéphane Chrétien [1]

[1] *Laboratoire ERIC, Université Lyon 2*
[2] *Halias Technologies*

Résumé. De récentes découvertes ont suggérées qu'utiliser l'ADN tumoral circulant (ctDNA) lors des tests sanguins procurait une meilleure précision que les méthodes standard pour le suivi des maladies cancéreuses. La ctDNA est relâchée dans l'organisme par le biais de plusieurs mécanisme biologique : apoptose, nécrose ou sécrétion active. Il est fortement envisagé que l'étude de la dynamique de cette ctDNA est la clé pour des alertes efficaces pour ce qui est de la détection précoce de cancer. De telles approches sont déjà utilisées dans l'industrie, GRAIL en étant un exemple.

Dans cet article, nous explorons l'utilisation de la théorie des Signatures pour la détection de tumeur cancéreuse à partir de relevés de ctDNA, modélisés par un processus stochastique bien choisi. La théorie des Signatures est un sujet dont l'intérêt grandissant est lié à ses remarquables performances. Cet outil réussi à capturer des dynamiques complexes par des quantificateurs géométriques non-linéaires. Sa possibilité d'estimation sur un faible échantillonnage est un atout important pour notre étude. Nous présentons ici une étude probabiliste des coefficients de signature d'un processus de mort avec immigration qui nous amène à considérer leur utilisation dans le cadre de la détection de tumeur maligne.

Mots-clés. Signature, Processus, Test d'hypothèse

Abstract. Recent discoveries have suggested that using circulating tumor DNA (ctDNA) that can be measured from blood samples provides more accuracy for cancer monitoring than standard methods. ctDNA potentially results from different mechanisms for DNA release from cells, such as apoptosis, necrosis and active secretion. It is strongly suspected that studying the dynamics of the ctDNA is key to producing alerts in the case of early cancer detection. Such approaches to cancer detection are being exploited in the industry, e.g. GRAIL in the UK.

In the present paper, we propose to explore the use of Signature theory for detecting aggressive cancerous tumors from blood samples using a model from the dynamics of ctDNA. Signatures are a topic of growing interest and can capture complex dynamics using non-linear features. They have moreover been proved very efficient for various machine learning challenges. One specific hurdle in designing such a detection method is the very small number of samples per patient. Our main discovery is that the theory of signatures can provide the tools for efficiently screening the patients when combined with a relevant birth and death process model for biomarker shedding of malignant tumors.

Keywords. Signature, Processes, Hypothesis testing

1 Introduction

La théorie des Signatures d'un chemin rugueux est un outil provenant de l'étude des chemins lisses par K.T. Chen [4]. En évaluant toutes les 1-formes différentielles créables à partir du chemin considéré sur le simplexe canonique on produit une suite infinie de quantificateurs géométriques permettant de caractériser chaque chemin de manière unique - à quelques invariances prêts. Fin 90, début 2000, T. Lyons, étends cet outil à la théorie des chemins rugueux. Depuis, l'utilisation des signatures a prouvé son efficacité sur plusieurs applications [8], la première ayant fait date restant l'interprétation de kanji chinois manuscrit.

De par sa nature désormais intrinsèquement liée aux chemins rugueux, de nombreuses études liées à la loi des signatures ont été menées[9]. Nous proposons ici un résumé d'un travail actuel dans lequel nous étudions la loi des premiers coefficients de signatures pour un processus de mort avec immigration, loi que nous utilisons par la suite pour établir un test d'hypothèse ayant pour but de distinguer dynamique bénigne (sous H_0) et maligne (sous H_1).

2 Prérequis sur les signatures et la dynamique de biomarqueurs cancéreux.

2.1 La théorie des Signatures.

La signature d'un chemin est un puissant outil d'extraction de caractéristiques géométriques pour un chemin multivarié. Elle est définie comme la suite infinie des intégrales itérées du chemin X : pour un chemin $X = (X^1, ..., X^d) : t \mapsto (X_t^1, ..., X_t^d)$ défini sur un intervalle $[a, b]$, et un multi-indice $i_1...i_k, \quad i_j \in \{1, ..., d\}$, on a

$$S_{[a,b]}^{i_1...i_k} = \int_{a<t_1<..<t_k<b} \cdots \int dX_{t_1}^{i_1} \dots dX_{t_k}^{i_k}.$$

La signature étant invariante par **reparamétrisation**, **translation** et **tree-like équivalence**, elle définie de manière unique le chemin X une fois celui ci considéré comme une classe d'équivalence pour les 3 relations ci-dessus.

Dans le cas des chemins définis par un processus stochastique, la signature est définie à l'aide des **intégrales d'Ito**, de manière à ce que l'on puisse caractériser la loi des coefficients de signatures.

Remarque : nous serons amené à étudier **l'aire de Levy** d'un chemin X, définie à partir des coefficients de signatures par la formule

$$\mathcal{LA}(X) = \frac{1}{2}\Big(S^{12}(X) - S^{21}(X)\Big).$$

2.2 Rudiments à propos de la dynamique des biomarqueurs cancéreux.

Les biomarqueurs cancéreux sont des brins d'ADN caractéristiques des cellules cancéreuses (ADN tumoral circulant : ctDNA) relachés principalement lors de **l'apoptose**, comme illustré dans la figure 1

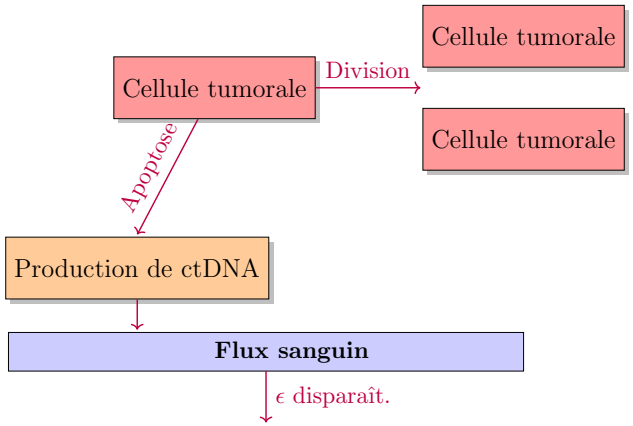


Figure 1: Processsus de réplication et d'apoptose pour des cellules tumorales : les brins de ctDNA sont relachés après apoptose avec une certaine probabilité q, puis sont détruit une fois dans le sang avec probabilité ϵ.

Le processus bénin : Les biomarqueurs tumoraux sont relâchés dans le sang avec une probabilité constante q, et disparaissent avec probabilité ϵ. En conséquence, nous modélisons ce phénomène par un **processus de mort pure avec immigration** C_t^B, avec un taux de mort ϵ, et un taux d'immigration λ_{bn} fixé par l'équation

$$A_0 \cdot \lambda_{\text{bn}} = A_0 \cdot d_{\text{bn}} \cdot q_d,$$

la population A_t étant considérée dans ce cas constante (en espérance) au cours du temps. La fonction génératrice des probabilités est donnée, pour ce type de processus, par :

$$\mathcal{C}^B(y,t) = \left(1 + (y-1)e^{-\varepsilon t}\right)^{C_0^B} e^{\frac{\lambda_{\text{bn}}}{\varepsilon} B_0 (y-1)\left(1-e^{-\varepsilon t}\right)}.$$

Quand t est grand, on a :

$$\lim_{t\to\infty} \mathcal{C}^B(y,t) = e^{\frac{\lambda_{\text{bn}}}{\varepsilon} B_0 (y-1)}.$$

indépendamment de la quantité initiale de biomarqueurs. On reconnait dans le membre de droite la fonction génératrice des probabilités d'une loi de Poisson de moyenne $B_0 \cdot \lambda_{\text{bn}}/\varepsilon$, et donc nous obtenons, en temps long :

$$C_t^B \sim \text{ Poisson } \left(\frac{B_0 \cdot \lambda_{\text{bn}}}{\varepsilon}\right).$$

Sans perte de généralité, nous pouvons considérer $\lambda_{\text{bn}} = \varepsilon$ et $B_0 = \mathbb{E}_{t\in\mathbb{R}_+}[C_t^B]$.

3 Tests d'hypothèses sur coefficients de signatures.

3.1 Conditions considérées.

On se place sous l'hypothèse (H_0) de **présence de biomarqueurs tumoraux provenant d'une tumeur bénigne**. Sous cette condition, nous émettons l'hypothèse que la quantité moyenne de ctDNA est constante.

De plus, dans l'optique de se placer au plus proche des conditions réelles, nous considérons que la première prise de sang intervient suffisamment longtemps après l'apparition de la tumeur pour supposer que le processus a convergé vers sa loi en temps long.

Nous considérons aussi que les temps d'échantillonnage sont assez parcimonieux pour supposer l'indépendance entre $C^B_{t_i}$ et $C^B_{t_{i-1}}$.

Pour s'adapter au cadre usuel de la théorie des Signatures, nous considérons les signaux $(t, \tilde{C^B_t})$ avec :

- $t \in [0, 1]$,
- $\tilde{C^B_t} = C^B_t - \mathbb{E}_{t\in[0,T]}[C^B_t]$

Par la suite, nous noterons $\text{Poisson}_c(0, \Lambda)$ la **loi Poisson**(Λ) **centrée** : si $Y \sim \text{Poisson}(\Lambda)$, et Y_c vérifie :

$$\mathbb{P}[Y_c = k - \Lambda] = \mathbb{P}[Y = k], \quad \forall k \in \mathbb{N}$$

alors $Y_c \sim \text{Poisson}_c(0, \Lambda)$.

Pour simplifier l'étude, on considèrera que $\Lambda = \frac{B_0.\lambda_{bn}}{\epsilon} = B_0 = \mu_{t\in[0,T]}(C^B_t)$

3.2 Variables pivots et test d'hypothèse.

Nous effectuons une procédure de tests multiples (voir Algorithme 1 sur les premiers coefficients de signature ainsi que **l'aire de Levy**, dont les lois peuvent être explicitées en utilisant les formules closes des coefficients et les hypothèses ci dessus.

4 Expériences

Pour évaluer la pertinence de ce test[1], nous l'avons évalué sur des réalisations de processus de vie et de mort. Les données testées sont formées de :

- 4000 courbes **bénignes**, $\lambda_{bn} \in \{0.1, 0.12, 0.14, 0.17\}$.

[1]Les expériences sont disponibles sur https://github.com/RemiVaucher/Thesis/tree/main/Chapter4/Oncology.

Algorithm 1 Test pour S^w avec $w \in \{2, 12, 21, 12-21\}$

Require: X et $w \in \{2, 12, 21, 12-21\}$ (avec la notation $12-21$ indiquant l'aire de Levy).
Ensure: (H_0) : la tumeur est bénigne.

Calcul de $\mu_C = \frac{1}{n}\sum_{i=0}^{n} C_{t_i}$ avec n le nombre d'observations de X

Calcul de $S^w(\tilde{X})$ avec $\tilde{X} = (t, C_t - \mu_C)$

Calcul de la p-valeur correspondant à la valeur observée de S^w avec $\Lambda = \mu_C$

Output: The selected tests for which H_0 is not corroborated by the observed data, using the mutliple testing correction proposed by, i.e. Benjamini-Hochberg [3].

- 36000 courbes avec **transition de régime bénin vers malin** à un temps aléatoire. Les taux de croissance malins appartiennent à l'ensemble

$$\mathcal{R} = \{0.001, 0.002, 0.004, 0.007, 0.01, 0.015\},$$

  et chaque taux r est scindé en 6 sous-groupes pour couvrir un maximum de paires (b, d).

- Dans tous les cas, les données sont **observées tous les 300 jours pour un total de 7 observations.**

- Pour compléter l'étude, nous avons aussi généré des processus purement malins.

Pour produire ces données, nous avons utilisé le package python BirDePy[6].

Les résultats de notre procédure sont présentés dans le Tableau 1

Table 1: Scores pour les différents tests.

Test	Accuracy	Precision	Recall	F1-score
T_2	0.934	0.996	0.999	0.998
T_{12}	0.934	0.996	0.996	0.996
T_{21}	0.934	0.996	0.999	0.998
T_{22}	0.934	0.996	0.999	0.998
T_{LA}	0.934	0.997	0.999	0.998
Test multiple corrigé	0.958	0.999	0.999	0.999

5 Conclusions et perspectives.

Au vu des résultats, notre procédure semble bien détecter les dynamiques malignes **sous réserves que ces dynamiques suivent correctement les processus stochastiques de vie et de mort de la littérature [5, 1].** Plusieurs pistes restent à exploiter :

- Étudier les limites de la procédure en espaçant encore les observations ou en repoussant le temps de transition au dernier intervalle inter-observations seulement.

- Un accès à de vraies données (pour l'instant difficile) permettrait d'évaluer la pertinence effective.
- Dans la perspective d'application à de vraie données médicales, il est envisagé d'utiliser les travaux sur la **prédiction conforme** pour obtenir des **p-valeurs conformes [7, 2].** Cette adaptation permettrait d'appliquer la procédure sans connaître au préalable la véritable loi des données.

References

[1] Stefano Avanzini, David M Kurtz, Jacob J Chabon, Everett J Moding, Sharon Seiko Hori, Sanjiv Sam Gambhir, Ash A Alizadeh, Maximilian Diehn, and Johannes G Reiter. A mathematical model of ctdna shedding predicts tumor detection size. *Science advances*, 6(50):eabc4308, 2020.

[2] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

[3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[4] Kuo-Tsai Chen. Integration of paths, geometric invariants and a generalized baker-hausdorff formula. *Annals of Mathematics*, 65(1):163–178, 1957.

[5] Richard Durrett and Richard Durrett. *Branching process models of cancer*. Springer, 2015.

[6] Sophie Hautphenne and Brendan Patch. Birth-and-death processes in python: The bird-epy package. *arXiv preprint arXiv:2110.05067*, 2021.

[7] Ying Jin and Emmanuel J Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

[8] Terry Lyons and Andrew D McLeod. Signature methods in machine learning. *arXiv preprint arXiv:2206.14674*, 2022.

[9] Terry Lyons and Hao Ni. Expected signature of brownian motion up to the first exit time from a bounded domain. 2015.

Développements théoriques pour l'apprentissage

Sur la non-convexité d'une contrainte de transport et ses conséquences en apprentissage machine

Lucas De Lara [1] & Mathis Deronzier [2] & Alberto González-Sanz [3] & Virgile Foy [4]

[1] *Université de Lorraine, France, lucas.de-lara@univ-lorraine.fr*
[2] *Université Paul Sabatier, France, mathis.deronzier@math.univ-toulouse.fr*
[3] *Columbia University, USA, ag4855@columbia.edu*
[4] *Université Paul Sabatier, France, virgile.foy@math.univ-toulouse.fr*

Résumé. L'opération de mesure image (aussi dite de transport) permet de redistribuer une mesure de probabilité via une fonction déterministe. Elle joue un rôle essentiel en statistiques et en optimisation : de nombreux problèmes d'apprentissage (notamment venant de l'équité algorithmique) incluent des contraintes ou des pénalités formulées comme des conditions de transport sur le modèle. Cependant, la littérature manque d'une appréciation théorique sur la (non) convexité de telles contraintes et ses conséquences sur les problèmes d'apprentissage associés. Cette contribution vise à combler cette lacune. Nous montrons que l'ensemble des fonctions produisant la même distribution de sortie à partir de deux distributions d'entrées distinctes est généralement non convexe. Puis, nous illustrons comment ce résultat induit des limitations strictes sur la conception de problèmes d'apprentissage équitable convexes.

Mots-clés. Mesure image, convexité, apprentissage machine, équité

Abstract. The push-forward operation enables one to redistribute a probability measure through a deterministic map. It plays a key role in statistics and optimization: many learning problems (notably algorithmic fairness) include constraints or penalties framed as push-forward conditions on the model. However, the literature lacks general theoretical insights on the (non)convexity of such constraints and its consequences on the associated learning problems. This paper aims at filling this gap. We prove that the set of maps inducing equal output distributions across distinct probability measures is generally nonconvex. Then, we show how this result implies critical limitations on the design of convex problems for learning fair predictors.

Keywords. Push-forward, convexity, machine learning, fairness

1 Introduction

Sachant une mesure de probabilité P sur $\mathbb{R}^d$ et une fonction mesurable $f : \mathbb{R}^d \to \mathbb{R}^p$, la *mesure image* de P par f est définie comme $f_\sharp P := P \circ f^{-1}$. Cette opération décrit la redistribution de la masse de P par l'allocation déterministe f, et joue un rôle croissant en statistiques et apprentissage machine. Notamment, plusieurs concepts d'équité de groupe peuvent se formuler comme trouver un modèle f tel que $f_\sharp P = f_\sharp Q$ (Dwork et al., 2012; Hardt et al., 2016).

Cependant, alors que la convexité est essentielle pour concevoir des problèmes d'apprentissage numériquement efficaces et statistiquement consistants (Hjort et Pollard, 1993; Bubeck et al., 2015), la littérature a peu étudié la (non) convexité de ces contraintes et ses conséquences.

Ce travail vise à combler cette lacune en étudiant l'ensemble des *fonctions égalisatrices* entre P et Q : les fonctions f telles que $f_\sharp P = f_\sharp Q$. Dans un premier temps, nous prouvons que cet ensemble est généralement non convexe. Dans un second temps, nous montrons comment cela rend généralement irréalisable la conception de problèmes d'apprentissage convexes impliquant des contraintes ou des pénalités de fonctions égalisatrices sur le modèle, tels que les problèmes d'apprentissage équitable. Ainsi, ce travail théorique a un intérêt pratique en fournissant des indications sur ce qui peut (ou ne peut pas) être réalisé en matière d'équité algorithmique.

2 Fonctions égalisatrices

Soit $\mathcal{G}$ l'ensemble des fonctions mesurables de $\mathbb{R}^d$ à $\mathbb{R}^p$, et $P, Q \in \mathcal{P}(\mathbb{R}^d)$ deux mesures de probabilité de Borel. Nous considérons l'ensemble des fonctions transformant P et Q en une même mesure arbitraire dans $\mathcal{P}(\mathbb{R}^p)$, que nous appelons les *fonctions égalisatrices* entre P et Q. Formellement, nous définissons

$$\mathcal{E}(P, Q) := \{f \in \mathcal{G} \mid f_\sharp P = f_\sharp Q\}.$$

La définition ci-dessus est motivée par des problèmes d'équité algorithmique, où l'on essaie typiquement de concevoir des modèles produisant les mêmes distributions de résultats pour des groupes protégés distincts. Commençons par remarquer que pour tout $P, Q \in \mathcal{P}(\mathbb{R}^d)$, $\mathcal{E}(P, Q) \neq \emptyset$ car il contient en particulier les fonctions constantes, et $\mathcal{E}(P, Q) = \mathcal{G}$ si $P = Q$.

Bien que nous ne fournissions pas une caractérisation complète de la (non) convexité de l'ensemble des fonctions égalisatrices, nous pouvons identifier des conditions suffisantes pour de larges classes de mesures de probabilité P et Q. La proposition ci-dessous détermine entièrement la convexité de $\mathcal{E}(P, Q)$ lorsque P, Q n'ont pas d'atome.

Proposition 2.1. *Soient $P, Q \in \mathcal{P}(\mathbb{R}^d)$ sans atomes telles que $P \neq Q$. Alors, $\mathcal{E}(P, Q)$ n'est pas convexe.*

De plus, dans le cas où P et Q sont à support fini, nous pouvons caractériser complètement la convexité de $\mathcal{E}(P, Q)$(De Lara et al., 2024a). Une implication de cette caractérisation concerne les mesures empiriques, comme le précise la proposition suivante.

Proposition 2.2. *Soient $n, m \geq 1$ deux entiers naturels, et deux ensembles $\{x_i\}_{i=1}^n, \{y_j\}_{j=1}^m \subset \mathbb{R}^d$ composés d'éléments distincts tels que $\{x_i\}_{i=1}^n \cap \{y_j\}_{j=1}^m = \emptyset$. Si $P := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ et $Q := \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$, alors:*

(i) si m et n sont premiers entre eux, alors $\mathcal{E}(P, Q)$ est l'ensemble des fonctions de $\mathcal{G}$ constantes $P + Q$-partout, qui est un ensemble convexe;

(ii) Si m et n ne sont pas premiers entre eux, alors $\mathcal{E}(P, Q)$ n'est pas convexe.

En résumé, Propositions 2.1 et 2.2 soulignent que la convexité est rare, et ne se produit parfois que parce que les fonctions égalisatrices sont constantes (c'est-à-dire triviales). Dans la suite du document, nous discutons des conséquences de ces résultats sur les problèmes d'apprentissage automatique.

3 Application à l'apprentissage

Cette section illustre le rôle de l'ensemble $\mathcal{E}(P, Q)$ dans des problèmes populaires d'apprentissage machine, et met en évidence les conséquences de sa (non) convexité. Á cette fin, nous introduisons d'abord un cadre générique pour étudier la convexité en apprentissage.

3.1 Apprentissage et convexité : un formalisme générique unifié

La plupart des problèmes d'apprentissage automatique consistent à minimiser un critère numérique (*e.g.*, la précision de la prédiction) sous une contrainte ou une pénalité (*e.g.*, la parcimonie). Formellement, supposons que $\mathcal{F} \subseteq \mathcal{G}$ soit un ensemble de *modèles de base* (*e.g.*, des réseaux de neurones avec une architecture fixe), et que $\mathcal{C} \subseteq \mathcal{G}$ soit une *contrainte*. Dans ce cas, $\mathcal{F} \cap \mathcal{C}$ représente l'ensemble des modèles *admissibles*. Dans les problèmes d'optimisation et d'apprentissage, une perte sert à quantifier l'écart d'un modèle par rapport à une certaine contrainte. Par souci de clarté, nous utilisons la terminologie suivante dans le reste du document.

Définition 3.1. *Soit* $\mathcal{C} \subseteq \mathcal{G}$. *Une fonction* $\mathrm{L} : \mathcal{G} \to [0, +\infty]$ *est une* $\mathcal{C}$-perte *si pour tout* $f \in \mathcal{G}$, $\mathrm{L}(f) = 0 \iff f \in \mathcal{C}$.

Pour une fonction de perte donnée $\mathrm{L} : \mathcal{G} \to [0, +\infty]$, un problème d'apprentissage incluant $\mathcal{C}$ dans son objectif correspond au problème d'optimisation *contraint*

$$\min_{f \in \mathcal{F} \cap \mathcal{C}} \mathrm{L}(f) \tag{1}$$

ou *pénalisé*, où $\lambda > 0$ gouverne un compromis entre L et une $\mathcal{C}$-perte $\mathrm{L}_{\mathcal{C}}$

$$\min_{f \in \mathcal{F}} \mathrm{L}(f) + \lambda \, \mathrm{L}_{\mathcal{C}}(f). \tag{2}$$

Un *problème de minimisation* est convexe si à la fois l'objectif à minimiser et l'ensemble des solutions admissibles sont convexes. Les chercheurs et les praticiens s'efforcent généralement de concevoir des problèmes d'optimisation convexes pour des raisons numériques (Bubeck et al., 2015) ou de consistance statistique (Haberman, 1989; Pollard, 1991; Niemiro, 1992; Arcones, 1998; Bartlett et al., 2006). Sur la base des formulations génériques ci-dessus, nous pouvons facilement identifier des conditions suffisantes pour la convexité dans les problèmes d'apprentissage. Problème 1 est convexe si $\mathcal{C}$ et $\mathcal{F}$ sont des ensembles convexes (car la convexité est stable sous intersection), et si L est une fonction convexe. Problème 2 est convexe si $\mathcal{F}$ est un ensemble convexe et si L et $\mathrm{L}_{\mathcal{C}}$ sont des fonctions convexes.

Soulignons le rôle spécifique de $\mathcal{C}$. Son influence est directe dans Problème 1, tandis qu'elle dépend d'une $\mathcal{C}$-perte $\mathrm{L}_{\mathcal{C}}$ dans Problème 2. Notamment, une $\mathcal{C}$-perte n'est évidemment pas déterminée de manière unique par $\mathcal{C}$, ce qui rend le choix de $\mathrm{L}_{\mathcal{C}}$ crucial pour atteindre la convexité dans Problème 2. Cependant, la conception d'une $\mathcal{C}$-perte convexe est parfois impossible, en conséquence d'un résultat classique de l'analyse convexe.

Théorème 3.2. *Si* $\mathrm{L} : \mathcal{G} \to [0, +\infty]$ *est une fonction convexe, alors* $\mathrm{L}^{-1}(\{0\})$ *est un ensemble convexe. Par conséquent, si* $\mathcal{C} \subseteq \mathcal{G}$ *est non convexe, alors il n'existe aucune* $\mathcal{C}$*-perte convexe.*

Cela fournit un critère simple pour vérifier si une condition $\mathcal{C}$ peut (ou ne peut pas) être quantifiée par une perte convexe : il suffit de vérifier que $\mathcal{C}$ lui-même est (ou n'est pas) un ensemble convexe. Les conséquences de ce résultat dans le domaine de l'apprentissage automatique sont importantes et peut-être mal appréciées : il peut être vain de chercher une pénalité convexe.

3.2 Fonctions égalisatrices et équité algorithmique de groupe

Nous passons maintenant des contraintes générales $\mathcal{C}$ aux cas spécifiques des fonctions égalisatrices. Si $\mathcal{C} := \mathcal{E}(P, Q)$, alors la convexité dépend de (P, Q) selon Section 2. Ainsi, de telles contraintes ne sont pas universellement convexes (c'est-à-dire convexes quelles que soient les mesures d'entrée) : elles sont même non convexes dans de nombreux scénarios classiques. En particulier, cela signifie par Théorème 3.2 qu'*il n'existe aucune perte* $\mathrm{L}_{\mathcal{E}(P,Q)}$ *qui soit convexe pour toutes les mesures* P *et* Q, contrairement (par exemple) à l'erreur quadratique moyenne qui est convexe quelle que soit la distribution des données. Il s'agit d'une limitation importante pour la conception de problèmes d'apprentissage convexes, puisque dans les scénarios typiques, les mesures sont des données d'entrée. Dans ce qui suit, nous montrons que ce cadre s'applique à l'équité algorithmique de groupe.

Soient $X : \Omega \to \mathbb{R}^{d-1}$ un vecteur aléatoire représentant certaines covariables, et $S : \Omega \to \{0, 1\}$ une variable aléatoire encodant un statut protégé binaire (par exemple, hommes et femmes) tel que $0 < \mathbb{P}(S = 1) < 1$. Une fonction $f \in \mathcal{G}$ satisfait *la parité statistique* par rapport à S si $f(X, S) \perp\!\!\!\perp S$ (Dwork et al., 2012). Nous définissons $P := \mathbb{L}((X, S) \mid S = 0)$ et $Q := \mathbb{L}((X, S) \mid S = 1)$ les mesures de probabilité conditionnelles des deux groupes protégés. Dans ce cas binaire, notons que la parité statistique peut être définie comme $\mathbb{L}(f(X, S) \mid S = 0) = \mathbb{L}(f(X, S) \mid S = 1)$, c'est-à-dire $f_\sharp P = f_\sharp Q$ donc $f \in \mathcal{E}(P, Q)$. Maximiser la précision du modèle sous la contrainte de parité statistique correspond donc à Problème 1 avec $\mathrm{L}(f) := \mathbb{E}\left[\|f(X, S) - Y\|^2\right]$ (ou toute autre perte convexe pour la précision) et $\mathcal{C} := \mathcal{E}(P, Q)$ (comme dans (Thibaut Le Gouic et al., 2020; Chzhen et al., 2020)). L'obtention d'un compromis entre précision et équité correspond à Problème 2 avec $\mathrm{L}_{\mathcal{C}}$ une perte de $\mathcal{E}(P, Q)$ (comme dans (Pérez-Suay et al., 2017; Risser et al., 2022)). Le plus souvent, on utilise $\mathrm{L}_{\mathcal{C}}(f) := D(f_\sharp P, f_\sharp Q)$ où $\mathcal{C}$ est (la puissance d') une distance ou une divergence.

La convexité d'un processus d'apprentissage impliquant la parité statistique a deux restrictions selon Section 3.1 : la convexité de $\mathcal{F}$ et celle de $\mathcal{E}(P, Q)$. Concernant la première : si $\mathcal{F}$ est un ensemble de réseaux de neurones d'architecture fixe, l'apprentissage n'est pas convexe ; si $\mathcal{F}$ est l'ensemble des fonctions linéaires, l'apprentissage pourrait être convexe

(mais probablement moins efficace sur des données complexes). Concernant la deuxième, la discussion de Section 3.1 et les résultats de Section 2 montrent que l'apprentissage est rarement convexe, que ce soit pour des mesures sans atome P, Q ou pour leurs équivalents empiriques. *Ainsi, aucun praticien ne peut concevoir un problème d'apprentissage par équité de groupe qui garantisse universellement la convexité.* Notamment, modifier $\mathrm{L}_{\mathcal{C}}$ (en utilisant par exemple un D convexe) est inutile.

L'approche la plus simple pour récupérer la convexité consiste à remplacer la contrainte $\mathcal{C}$ par une variante convexe. Par exemple, plutôt que d'exiger que $f(X, S) \perp\!\!\!\perp S$, on pourrait simplement exiger que $\mathbb{C}\mathrm{ov}(f(X, S), S) = 0$ où $\mathbb{C}\mathrm{ov}$ est la covariance entre deux variables aléatoires. C'est notamment ce que Zafar et al. (2017, 2019) ont fait pour obtenir une relaxation convexe de la parité statistique dans leurs cadres de classification équitable. Inversement, on peut aussi atteindre la convexité en renforçant la contrainte. Contrairement à la parité statistique, qui est une définition distributionnelle (ou de groupe) de l'équité, *l'équité contrefactuelle* se concentre sur les prédictions au niveau des entrées Kusner et al. (2017). Un aspect intéressant exploré dans Kusner et al. (2017); De Lara et al. (2024b) réside dans le fait que (sous des hypothèses standards) elle implique la parité statistique en étant sa contrepartie à l'échelle des entrées. Surtout, l'équité contrefactuelle représente une contrainte convexe qui peut être quantifiée par une perte convexe Russell et al. (2017); De Lara et al. (2024b). On peut donc l'utiliser comme une restriction convexe de la parité statistique. Cependant, cela soulève d'autres défis, comme le calcul du modèle causal nécessaire à l'équité contrefactuelle.

Bibliographie

M. A. Arcones. Asymptotic theory for M-estimators over a convex kernel. *Econometric Theory*, 14(4):387–422, 1998.

P. L. Bartlett, M. I. Jordan, et J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, et M. Pontil. Fair regression with Wasserstein barycenters. In *Advances in Neural Information Processing Systems*, volume 33, pages 7321–7331. Curran Associates, Inc., 2020.

L. De Lara, M. Deronzier, A. González-Sanz, et V. Foy. On the nonconvexity of push-forward constraints and its consequences in machine learning. *arXiv preprint arXiv:2403.07471*, 2024a.

L. De Lara, A. González-Sanz, N. Asher, L. Risser, et J.-M. Loubes. Transport-based counterfactual models. *Journal of Machine Learning Research*, 25(136):1–59, 2024b.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, et R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

S. J. Haberman. Concavity and estimation. *The Annals of Statistics*, pages 1631–1661, 1989.

M. Hardt, E. Price, et N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

N. L. Hjort et D. Pollard. Asymptotics for minimisers of convex processes. *arXiv preprint arXiv:1107.3806*, 1993.

M. J. Kusner, J. Loftus, C. Russell, et R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076. Curran Associates, Inc., 2017.

W. Niemiro. Asymptotics for M-estimators defined by convex minimization. *The Annals of Statistics*, pages 1514–1533, 1992.

A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, et G. Camps-Valls. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 339–355. Springer, 2017.

D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.

L. Risser, A. Gonzalez-Sanz, Q. Vincenot, et J.-M. Loubes. Tackling algorithmic bias in neural-network classifiers using Wasserstein-2 regularization. *Journal of Mathematical Imaging and Vision*, pages 1–18, 2022.

C. Russell, M. J. Kusner, J. Loftus, et R. Silva. When worlds collide: Integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Thibaut Le Gouic, Jean-Michel Loubes, et Philippe Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

M. B. Zafar, I. Valera, M. G. Rogriguez, et K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *International Conference on Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, et K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

General reproducing properties in RKHS with application to derivative operator

Fatima-Zahrae EL-BOUKKOURI* Josselin GARNIER†
Olivier ROUSTANT‡

Abstract

Dans cet article, nous généralisons la propriété reproduisante dans les espaces de Hilbert à noyaux reproduisants (RKHS). Nous établissons une propriété de reproduction pour l'adhérence de la classe des combinaisons d'opérateurs de composition sous des conditions minimales. Comme application, nous améliorons la condition suffisante connue jusqu'alors, pour que la propriété de reproduction soit valable pour l'opérateur de dérivation. Ces résultats permettent d'élargir le cadre d'application du théorème du représentant pour les algorithmes d'apprentissage régularisés qui impliquent des données sur les valeurs de fonctions, les gradients ou tout autre opérateur de la classe considérée.

Abstract

In this paper, we generalize the reproducing property in Reproducing Kernel Hilbert Spaces (RKHS). We establish a reproducing property for the closure of the class of combinations of composition operators under minimal conditions. As an application, we improve the existing sufficient condition for the reproducing property to hold for the derivative operator. These results extend the scope of applicability of the representer theorem for regularized learning algorithms that involve data for function values, gradients, or any other operator from the considered class.

Keywords: Reproducing kernel Hilbert spaces, Reproducing property, Representer theorem, Combination of composition operators, Derivative operator.

Introduction

Machine learning algorithms often involve penalized regression problems of the form

$$\min_{h\in\mathcal{H}} \sum_{i=1}^{n} (Lh(x_i) - y_i)^2 + \lambda \|h\|^2.$$

*Institut de Mathématiques de Toulouse, Université de Toulouse, INSA, Toulouse, 31077 cedex 4, France, (el-boukkouri@insa-toulouse.fr)

†CMAP, CNRS, Ecole polytechnique, Institut Polytechnique de Paris, Palaiseau, 91120, France, josselin.garnier@polytechnique.edu

‡Institut de Mathématiques de Toulouse, Université de Toulouse, INSA, Toulouse, 31077 cedex 4, France, (roustant@insa-toulouse.fr)

Here $(x_i, y_i)_{1 \leq i \leq n}$ is a dataset where x_i belongs to some set $\mathbb{X}$ and $y_i \in \mathbb{R}$, $\lambda \in \mathbb{R}_+$ is a penalty coefficient, $\mathcal{H}$ is a Hilbert space of real-valued functions on $\mathbb{X}$, and L is a linear operator on $\mathcal{H}$. When $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), it turns out that the solution lives in a finite-dimensional space, a famous result known as "representer theorem" [1, 2].

A key property to prove this result is the generalized "reproducing property"

$$\forall x \in \mathbb{X}, \forall h \in \mathcal{H} : \qquad Lh(x) = \langle h, \widetilde{L}(x) \rangle \tag{1}$$

where $\widetilde{L}(x)$ belongs to $\mathcal{H}$. When L is the identity operator, we have $\widetilde{L}(x) = K(x, .)$ where K is the kernel associated to $\mathcal{H}$,

$$\forall x \in \mathbb{X}, \forall h \in \mathcal{H} : \qquad h(x) = \langle h, K(x, .) \rangle$$

which is the original reproducing property of RKHS [1]. This immediately extends to finite linear combinations, i.e. when L has the form

$$Lf(x) = \sum_{i=1}^{q} \alpha_i f(v_i(x))$$

where $\alpha_i \in \mathbb{R}$ and $v_i(x) \in \mathbb{X}$. In that case, $\widetilde{L}(x) = \sum_{i=1}^{q} \alpha_i K(v_i(x), .)$. However, the generalization to more complex operators such as derivative or integral operators is not straightforward as it involves a passage to the limit.

In this paper, we consider a broad class of operators corresponding to limits of linear combinations and we give a necessary and sufficient condition on K for (1) to hold under the assumption that (1) remains true when passing to the limit (Section 1). Then we focus on the derivative operator, and show that the generalized reproducing property holds if the cross derivative of the kernel exists and is continuous on the diagonal of $\mathbb{X} \times \mathbb{X}$. We prove that this condition is less restrictive than the known sufficient condition that K is of class C^2 [3, 4] (Section 2).

1 Reproducing property for linear operators

A RKHS $\mathcal{H}$ is a Hilbert space of real-valued functions defined on a set $\mathbb{X}$ such that for all $x \in \mathbb{X}$, the evaluation $h \in \mathcal{H} \mapsto h(x)$ is continuous. Giving a RKHS is equivalent to giving a positive semidefinite function, or kernel, K. The link between $\mathcal{H}$ and K is given by the reproducing property:

$$\forall f \in \mathcal{H}, \forall x \in \mathbb{X} : \qquad f(x) = \langle f, K(\cdot, x) \rangle_{\mathcal{H}}$$

We first recall the definition of so-called combination of composition operators, introduced in [5].

Definition 1. *Let $v : \mathbb{X} \to \mathbb{X}$ be an arbitrary function. The composition operator T_v, with function v, is defined by:*

$$T_v : \mathcal{H} \to \mathcal{F}(\mathbb{X}, \mathbb{R}), \quad T_v(f) := f \circ v,$$

We call a combination of composition operators with functions v_i and weights $\alpha_i \in \mathbb{R}$ for $1 \leq i \leq q$ the operator:

$$L = \sum_{i=1}^{q} \alpha_i T_{v_i}$$

The class of combinations of composition operators on functions on $\mathcal{H}$ is denoted $\mathcal{CC}(\mathcal{H})$.

Let $L = \sum_{i=1}^{q} \alpha_i T_{v_i} \in \mathcal{CC}(\mathcal{H})$. Notice that for any $x \in \mathbb{X}$, $Lf(x) = \sum_{i=1}^{q} \alpha_i f(v_i(x))$. Thus, the linear form $f \in \mathcal{H} \mapsto Lf(x)$ is continuous as a linear combination of evaluations of $\mathcal{H}$. Following [6, §4.4.], we will denote by $\widetilde{L}(x) \in \mathcal{H}$ its representer, i.e. the unique element of $\mathcal{H}$ verifying

$$Lf(x) = \langle f, \widetilde{L}(x) \rangle \tag{2}$$

We have explicitly

$$\widetilde{L}(x) = \sum_{i=1}^{q} \alpha_i K(v_i(x), .). \tag{3}$$

The class $\mathcal{CC}(\mathcal{H})$ is limited to finite linear combinations. We now define a natural extension of this class by considering its closure with respect to pointwise convergence.

Definition 2. *We define the class of operators*

$$\mathcal{C}(\mathcal{H}) = \left\{ L : \mathcal{H} \to \mathcal{F}(\mathbb{X}, \mathbb{R}) : \exists (L_n)_{n \in \mathbb{N}} \in \mathcal{CC}(\mathcal{H})^{\mathbb{N}} \text{ s.t. } L(f) = \lim_{n \to +\infty} L_n(f) \, \forall f \in \mathcal{H} \right\}$$

where the limit is in the pointwise sense.

If $L \in \mathcal{C}(\mathcal{H})$, writing $L_n = \sum_{i=1}^{q_n} \alpha_{i,n} T_{v_{i,n}}$ (where $q_n \in \mathbb{N}, \alpha_{i,n} \in \mathbb{R}, v_{i,n} \in \mathcal{F}(\mathbb{X}, \mathbb{X})$), we have for all $f \in \mathcal{H}$ and all $x \in \mathbb{X}$,

$$L(f)(x) = \lim_{n \to +\infty} \sum_{i=1}^{q_n} \alpha_{i,n} f(v_{i,n}(x))$$

As an example, when $\mathbb{X} = \mathbb{R}$, the class $\mathcal{C}(\mathcal{H})$ includes derivative operator. In fact, if the functions of $\mathcal{H}$ are differentiable, the derivative operator $L : f \mapsto f'$ is written as $\lim_{n \to +\infty} L_n$ where $L_n = nT_{v_{2,n}} - nT_{v_{1,n}}$ with $v_{2,n} : x \mapsto x + \frac{1}{n}$ and $v_{1,n} : x \mapsto x$. In that case, $\widetilde{L_n}(x) = nK(v_{2,n}(x), .) - nK(v_{1,n}(x), .)$ for all $x \in \mathbb{R}$.
For any operator L and for any function $F : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ denote

$$L^{\ell} F : \mathbb{X} \times \mathbb{X} \to \mathbb{R}, \quad (x, y) \mapsto L(F(\cdot, y))(x),$$

and

$$L^{r} F : \mathbb{X} \times \mathbb{X} \to \mathbb{R}, \quad (x, y) \mapsto L(F(x, \cdot))(y).$$

Notice that with these notations, for any $x, y \in \mathbb{X}$ we have $(L^{\ell} F)(., y) = L(F(., y))$ and $(L^{r} F)(x, .) = L(F(x, .))$. We will also simply write $L^{\ell} L^{r}$ instead of $L^{\ell} o L^{r}$.

Theorem 1. *Let $\mathcal{H}$ be a RKHS with kernel K. Let $(L_n)_{n \in \mathbb{N}}$ be a sequence of $\mathcal{CC}(\mathcal{H})$. We denote by $\widetilde{L_n}(x) \in \mathcal{H}$ the representer of the linear form $f \in \mathcal{H} \mapsto L_n f(x)$ (see (2)). The following conditions are equivalent:*

(i) $(L_n^\ell L_m^r K(x,x))_{n,m}$ *converges when* n, m *tend independently to* $+\infty$ *for all* $x \in \mathbb{X}$.

(ii) $(\widetilde{L_n}(x))_{n\in\mathbb{N}}$ *converges in* $\mathcal{H}$ *for all* $x \in \mathbb{X}$.

In that case, denote $\widetilde{L}(x) := \lim_{n\to+\infty} \widetilde{L_n}(x) \in \mathcal{H}$. *Then, for all* $f \in \mathcal{H}$, $L_n f$ *converges pointwise. Define the operator* $L \in \mathcal{C}(\mathcal{H})$ *by* $Lf(x) = \lim_{n\to+\infty} L_n f(x)$ *(for all* $f \in \mathcal{H}$, $x \in \mathbb{X}$*). Then, for all* $x \in \mathbb{X}$ *the mapping* $f \in \mathcal{H} \mapsto Lf(x)$ *is continuous and the reproducing property holds:*

$$\forall x \in \mathbb{X}, \forall f \in \mathcal{H} : \qquad Lf(x) = \langle f, \widetilde{L}(x) \rangle$$

Finally, $\|\widetilde{L}(x)\|^2 = L^\ell L^r K(x,x)$.

2 Derivative Reproducing Property

The reproducing property for the derivative operator has been extensively studied in the literature under various conditions. For example, in [3], this result is established for Mercer kernels of class C^2. However, the reproducing property holds, as we will demonstrate, under a broader condition, where the kernel is not required to be C^2.

Corollary 1. *Let* $\mathcal{H}$ *be a RKHS of real-valued functions defined on a non-empty open set* $\mathbb{X} \subset \mathbb{R}$, *with reproducing kernel* K. *If* K *is of class* C^1 *and* $\frac{\partial^2 K}{\partial x_1 \partial x_2}$ *exists and is continuous on a neighborhood of the diagonal, then for any* $x \in \mathbb{X}$,

$$\frac{K(x+t,\cdot) - K(x,\cdot)}{t}$$

converges in $\mathcal{H}$ *to* $\frac{\partial K}{\partial x_1}(x,.)$ *as* t *tends to* 0. *Furthermore, all functions of* $\mathcal{H}$ *are differentiable, the mapping* $f \in \mathcal{H} \mapsto f'(x)$ *is continuous, and the derivative reproducing property holds:*

$$\forall x \in \mathbb{X}, \forall f \in \mathcal{H} : \qquad f'(x) = \langle f, \frac{\partial K}{\partial x_1}(x,.) \rangle$$

Finally, $\|\frac{\partial K}{\partial x_1}(x,.)\|^2 = \frac{\partial^2 K}{\partial x_1 \partial x_2}(x,x)$.

Corollary 1 gives a sufficient condition on K for a derivative reproducing property to hold, namely the existence of the cross-derivative and its continuity on a neighborhood of the diagonal. This is less restrictive than the C^2 condition on K found in the literature. Indeed, the following example exhibits a kernel K that is not of class C^2 but such that $\frac{\partial^2 K}{\partial x_1 \partial x_2}$ exists and is continuous.

Example 1 (Non-C^2 Kernel with finite cross derivative)**.** *Consider the function:*

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = \begin{cases} x^4 \sin\left(\frac{1}{x}\right) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

We notice that f *is twice differentiable on* $\mathbb{R}$. *For* $x \in \mathbb{R}^*$ *, the first and second derivatives are:*

$$f'(x) = \begin{cases} 4x^3 \sin\left(\frac{1}{x}\right) - x^2 \cos\left(\frac{1}{x}\right)) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

and,

$$f''(x) = \begin{cases} 12x^2 \sin\left(\frac{1}{x}\right) - 6x\cos\left(\frac{1}{x}\right) - \sin\left(\frac{1}{x}\right) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Notice that f'' has no limit at 0, due to the term $\sin(\frac{1}{x})$.
Consider now the rank-one kernel K:

$$K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \quad K(x,y) = f(x)f(y).$$

K is not of class C^2 since for all $y \in \mathbb{R}$, the function $x \mapsto \frac{\partial^2 K}{\partial x_1^2}(x,y) = f''(x)f(y)$ is not continuous at $x = 0$. Nevertheless, the cross derivative $\frac{\partial^2 K}{\partial x_1 \partial x_2} = f'(x)f'(y)$ exists for all $x, y \in \mathbb{R}$ and is continuous.

References

[1] N. Aronszajn. "Theory of Reproducing Kernels". In: *Transactions of the American Mathematical Society* 68 (1950), pp. 337–404.

[2] R. Herbrich B. Schölkopf and Alexander J. Smola. "A Generalized Representer Theorem". In: Lecture Notes in Computer Science 2111 (2001). Ed. by David Helmbold and Ben Williamson, pp. 416–426.

[3] Ding-Xuan Zhou. "Derivative reproducing properties for kernel methods in learning theory". In: *Journal of Computational and Applied Mathematics* (2007). Received 27 June 2007.

[4] J. C. Ferreira and V. A. Menegatto. "Positive definiteness, reproducing kernel Hilbert spaces and beyond". In: *Ann. Funct. Anal.* 4.1 (2013), pp. 64–88. ISSN: 2008-8752.

[5] David Ginsbourger, Nicolas Durrande, and Olivier Roustant. "Kernels and designs for modelling invariant functions: From group invariance to additivity". In: *MODA-10*. Physica-Verlag HD, 2013.

[6] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

On Volume Minimization in Conformal Regression

Batiste Le Bars [1] & Pierre Humbert [2]

[1] *Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189, CRIStAL, F-59000 Lille*
batiste.le-bars@inria.fr
[2] *Sorbonne Université et Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France*
pierre.humbert@sorbonne-universite.fr

Abstract. We study the question of volume optimality in split conformal regression, a topic still poorly understood in comparison to coverage control. Using the fact that the calibration step can be seen as an empirical volume minimization problem, we first derive a finite-sample upper-bound on the excess volume loss of the interval returned by the classical split method. This important quantity measures the difference in length between the interval obtained with the split method and the shortest oracle prediction interval. Then, we introduce `EffOrt`, a methodology that modifies the learning step so that the base prediction function is selected in order to minimize the length of the returned intervals. In particular, our theoretical analysis of the excess volume loss of the prediction sets produced by `EffOrt` reveals the links between the learning and calibration steps, and notably the impact of the choice of the function class of the base predictor.

Keywords. Conformal Prediction, Minimum Volume Sets

1 Introduction

Conformal Prediction (CP) (Vovk et al., 2005) has recently been considered as one of the state-of-art technique to construct distribution-free prediction sets satisfying probabilistic coverage guarantees. Formally, consider a random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ and some coverage level $\alpha \in [0, 1]$, CP techniques construct a set-valued function $C : \mathcal{X} \to 2^{\mathcal{Y}}$ such that:

$$\mathbb{P}(Y \in C(X)) \geq 1 - \alpha \, . \tag{1}$$

This is particularly useful when the user prefers to be confident with the range of values that Y can take, rather than having only a single predicted scalar value. While these techniques are completely distribution-free, making them quite powerful in practice, they still suffer from an important limitation: how can we be sure that the trivial prediction set $C(x) = \mathcal{Y}$ is not returned? The CP literature therefore suggests to not only satisfy the coverage, but also look at the size of the predicted sets. The smaller is a prediction set, the more *efficient* it is considered. However, most works do this analysis empirically, while very few has been focusing on the statistical control of the size of CP sets.

In this paper, we therefore propose to study when $C(x)$ is in fact a solution of an optimization problem of the form:

$$\min_{C} \ \mathbb{E}[\mu(C(X))] \quad \text{s.t.} \quad \mathbb{P}(Y \in C(X)) \geq 1 - \alpha \, , \tag{2}$$

where μ is a measure of the volume of the set $C(x)$, typically the Lebesgue measure in regression problems, or the counting measure in classification. This optimization problem ensures that among all prediction sets $C(x)$ that satisfy the coverage condition (1), the volume of the returned set is also of minimal size. Looking at problem (2) instead of (1) alone is therefore more meaningful as it encapsulates the two key aspects of CP: coverage and efficiency.

2 Background

2.1 Preliminaries on Split Conformal Prediction

In this section, we give some important reminders on CP, focusing on the split approach at the core of this paper (Papadopoulos et al., 2002).

Let us assume that we have access to a data set $\mathcal{D} = \{(X_i, Y_i)\}_{1 \leq i \leq n}$, that we split into a learning set $\mathcal{D}^{lrn}$ and a non-overlapping calibration set $\mathcal{D}^{cal}$, containing respectively $n_\ell \geq 1$ and $n_c \geq 1$ data points such that $n_\ell + n_c = n$.

The first step of split CP, referred to as the *learning step*, consists in finding a *base* predictor $f \in \mathcal{F}$ using the learning data set $\mathcal{D}^{lrn}$. This predictor, denoted $\hat{f}$, is then used to define a nonconformity score function $s = s_{\hat{f}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, such that for a pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $s_{\hat{f}}(x, y)$ measures the level of non-conformity of the point (x, y) with respect to the base predictor $\hat{f}$. In other word, it measures how far is the true value y from the prediction $\hat{f}(x)$. Whether we are in the regression or classification setting, many possible base predictors and score functions exist in the literature (see e.g. Angelopoulos and Bates (2023)).

In the second step of split CP, referred to as the *calibration step*, we construct the prediction set. To this end, we first calculate the values of $s_{\hat{f}}$ taken on the calibration set $\mathcal{D}^{cal}$, called the nonconformity scores $S_i := s_{\hat{f}}(X_i, Y_i)$, $i \in [\![n_c]\!]$. Then, we compute the $\lceil (n_c + 1)(1 - \alpha) \rceil$-th smallest nonconformity score $\hat{q}_{1-\alpha} := S_{(\lceil (n_c+1)(1-\alpha) \rceil)}$, where $S_{(1)} \leq \ldots \leq S_{(n_c)}$, and we return, for any $x \in \mathcal{X}$, the set-valued function $\widehat{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ such that $\forall x \in \mathcal{X}$:

$$\widehat{C}(x) := \left\{ y \in \mathcal{Y} \, : \, s_{\hat{f}}(x, y) \leq \hat{q}_{1-\alpha} \right\}. \tag{3}$$

Quite remarkably, if we only assume that the scores $S_1, \ldots, S_{n_c}$ and $s(X, Y)$ are *exchangeable*, the set (3) satisfies condition (1) (Papadopoulos et al., 2002).

2.2 Problem statement

In this work, we focus on conformal regression problems with $\mathcal{Y} = \mathbb{R}$. Precisely, we study when and how split CP outputs prediction sets approximating the solution of Problem (2). Since we consider regression tasks, let us first re-write the latter optimization problem by replacing μ with the Lebesgue measure $\lambda : \mathcal{B}(\mathbb{R}) \to [0, +\infty]$, $\mathcal{B}(\mathbb{R})$ being the Borel σ-algebra on $\mathbb{R}$:

$$\min_{C \in \mathcal{C}_{\text{Borel}}} \mathbb{E}[\lambda(C(X))] \text{ s.t. } \mathbb{P}(Y \in C(X)) \geq 1 - \alpha, \tag{4}$$

where $\mathcal{C}_{\text{Borel}} := \{\text{Measurable functions } C : \mathcal{X} \to \mathcal{B}(\mathbb{R})\}$. In the following, we refer to C^* as one minimizer of (4). Note that optimizing over all possible measurable functions in $\mathcal{C}_{\text{Borel}}$ can be difficult in practice and in this work, we will restrict the space of research $\mathcal{C}_{\text{Borel}}$ in (4) to smaller sets of set-valued functions, namely those outputting intervals. Like in statistical learning theory, this restriction can be thought of as a source of *approximation error*. In other words, we would like the restricted set to be sufficiently complex so that it includes (one of) the function solving (4). If it is not the case, we face such an approximation error. Nevertheless, controlling this error is not the objective of this paper, as we are going to mostly focus on the *estimation error*, which comes from the fact that only an empirical version of (4) is going to be solved.

3 Restriction to intervals with constant size

In this section, we restrict the space of research in Problem (4) to the class of prediction sets $\mathcal{C}_{\mathcal{F}}^{\text{const}} = \{C_{f,t}(\cdot) = [f(\cdot) - t, f(\cdot) + t]; f \in \mathcal{F}, t \geq 0\}$. Notice that for simplicity of exposition, and because it does not depend on x, in this section the expected size of $C_{f,t} \in \mathcal{C}_{\mathcal{F}}^{\text{const}}$ is simply denoted $\lambda(C_{f,t}) = 2t$.

3.1 Base predictor $f \in \mathcal{F}$ is given: optimality of the conformal step

We first start in the setting where the base predictor f is given, meaning that we do not consider the learning phase. Over $\mathcal{C}_{\mathcal{F}}^{\text{const}}$, the optimization problem (4) becomes:

$$\min_{t \geq 0} \; 2t \text{ s.t. } \mathbb{P}(|Y - f(X)| \leq t) \geq 1 - \alpha \,. \tag{5}$$

Denoting by $S = |Y - f(X)|$ the random variable of the absolute residual, the solution of the above optimization corresponds to the quantile of order $1 - \alpha$ of the random variable S. More formally, if we denote by $Q(\,\cdot\,; S) : [0, 1] \to \mathbb{R}$ the quantile function of S, then the optimal value solving (5) is exactly $t^* = Q(1 - \alpha; S)$ and the associated optimal set is $C_{f,t^*}^{1-\alpha}(x) = [f(x) - t^*, f(x) + t^*]$.

Importantly, notice that the conformal step of the original split CP in fact solves an empirical version of the previous problem, but with a slightly increased coverage:

$$\min_{t \geq 0} \quad 2t \text{ s.t. } \frac{1}{n_c} \sum_{i=1}^{n_c} \mathbf{1}\{|Y_i - f(X_i)| \leq t\} \geq \frac{(1 - \alpha)(n_c + 1)}{n_c} \,, \tag{6}$$

with solution $\hat{t} = S_{(\lceil (n_c+1)(1-\alpha) \rceil)}$ and associated set denoted $C_{f,\hat{t}}^{1-\alpha}(x) = [f(x) - \hat{t}, f(x) + \hat{t}]$. As mentioned above, $\hat{t}$ is the quantity computed during the calibration step of the split CP method (see Section 2.1). It corresponds to the empirical quantile function of S, defined by $\widehat{Q}(q; \{S_i\}_{i=1}^{n_c}) := \inf\{t : \frac{1}{n_c} \sum_{i=1}^{n_c} \mathbf{1}\{S_i \leq t\} \geq q\}$, evaluated at $(1 - \alpha)(n_c + 1)/n_c$.

Let us now study the excess volume loss of $C_{f,\hat{t}}^{1-\alpha}$ which is measured by the difference in length between $C_{f,t^*}^{1-\alpha}$ and $C_{f,\hat{t}}^{1-\alpha}$. To obtain a finite-sample upper bound on this difference, we must

consider some regularity assumption on the distribution of S, and more particularly on its quantile function.

Assumption 1. (Regularity condition). *Let $S = |Y - f(X)|$. $\forall f \in \mathcal{F}$, $\forall \alpha \in (0,1), \exists r, \gamma \in (0,1]$ and $L > 0$ such that $Q(\cdot; S)$ is locally (γ, L)-Hölder continuous, i.e. $\forall q_1, q_2 \in [1-\alpha-r, 1-\alpha+r]$:*

$$|Q(q_1; S) - Q(q_2; S)| \leq L|q_1 - q_2|^\gamma \, .$$

Proposition 1. *Let Assumption 1 and let $\hat{t} = \widehat{Q}((1-\alpha)\frac{n_c+1}{n_c}; \{S_i\}_{i=1}^{n_c})$ and $C_{f,\hat{t}}^{1-\alpha}$ the corresponding set. If the points in $\mathcal{D}^{cal}$ are i.i.d., if $(n_c+1)(1-\alpha)$ is not an integer, and if n_c is large enough so that $\frac{1-\alpha}{n_c} + \sqrt{\frac{\log(2/\delta)}{2n_c}} \leq r$, then with probability greater than $1-\delta$:*

$$\lambda\Big(C_{f,\hat{t}}^{1-\alpha}\Big) \leq \lambda\Big(C_{f,t^*}^{1-\alpha}\Big) + 2L\Big(\frac{1}{n_c} + \sqrt{\frac{\log(2/\delta)}{2n_c}}\Big)^\gamma \, . \tag{7}$$

The previous result provides an excess volume upper-bound for $C_{f,\hat{t}}^{1-\alpha}$ compared to the oracle $C_{f,t^*}^{1-\alpha}$. This bound does not only confirm the asymptotic optimality of the conformal procedure when f is given, but also provides a rate of convergence dominated by $\tilde{\mathcal{O}}(n_c^{-\gamma})$ when we get rid of constants and log factors.

3.2 Base predictor $f \in \mathcal{F}$ is *not* given: sub-optimality of the least-square regressor

Now, we investigate how f should be learned during the learning step to obtain a prediction interval of minimal size. Let us consider Problem (4) over $\mathcal{C}_{\mathcal{F}}^{\text{const}}$: $\min_{f \in \mathcal{F}, t \geq 0} \; 2t$ s.t. $\mathbb{P}(|Y - f(X)| \leq t) \geq 1-\alpha$. By replacing t with its optimal value as a function of f, i.e. $t^* = Q(1-\alpha; |Y - f(X)|)$, we obtain what we call the $(1-\alpha)$-QAE problem (Quantile Absolute Error):

$$\min_{f \in \mathcal{F}} Q(1-\alpha; |Y - f(X)|) \, . \tag{8}$$

In words, this optimization problem tells us that f should minimize the $(1-\alpha)$-quantile of the distribution of $S = |Y - f(X)|$. This is quite natural, since this quantile is the one selected to build the prediction interval, and the smaller it is, the smaller the interval will be.

3.2.1 `EffOrt`: EFFiciency-ORienTed split conformal regression

In this section, we propose a methodology to approach the oracle prediction set $C_{f^*,t^*}^{1-\alpha}(x) = [f^*(x) - t^*, f^*(x) + t^*]$, with f^* the minimizer of the $(1-\alpha)$-QAE (Problem (8)) and $t^* = Q(1-\alpha; |Y - f^*(X)|)$. We place ourselves in the split conformal framework of Section 2.1, having access to a learning data set $\mathcal{D}^{lrn}$ used to learn f, and a calibration data set $\mathcal{D}^{cal}$. With a slight abuse of notation we will write $i \in \mathcal{D}^{lrn}$ or $\mathcal{D}^{cal}$ to indicate $(X_i, Y_i) \in \mathcal{D}^{lrn}$or $\mathcal{D}^{cal}$.

The proposed methodology, referred to as `EffOrt`, consists in the following steps:

3.2 Base predictor $f \in \mathcal{F}$ is not given: sub-optimality of the least-square regressor

1. Learn $\hat{f} \in \underset{f \in \mathcal{F}}{\arg\min}\, \widehat{Q}(1-\alpha; \{|Y_i - f(X_i)|\}_{i \in \mathcal{D}^{lrn}})$, i.e. minimize the empirical version of the $(1-\alpha)$-QAE following the procedure of Pena-Ordieres et al. (2020)
2. Proceed to the calibration step, i.e. take $\hat{t} = \widehat{Q}\Big((1-\alpha)\frac{n_c+1}{n_c}; \{|Y_i - \hat{f}(X_i)|\}_{i \in \mathcal{D}^{cal}}\Big)$
3. For any test point $X \in \mathcal{X}$, output the prediction interval $C^{1-\alpha}_{\hat{f},\hat{t}}(X) = [\hat{f}(X) - \hat{t}, \hat{f}(X) + \hat{t}]$

3.2.2 Theoretical analysis

In this last subsection, we theoretically analyze the performance of the prediction set output by `EffOrt`. To this aim, we require the following assumption.

Assumption 2. *There exists $\phi(\mathcal{F}, \delta, n) < +\infty$ such that with probability at least $1-\delta$:*

$$\sup_{\substack{t \geq 0 \\ f \in \mathcal{F}}} \Big| \mathbb{P}\left(|Y - f(X)| \leq t\right) - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{|Y_i - f(X_i)| \leq t\} \Big| \leq \phi(\mathcal{F}, \delta, n)\ .$$

In this assumption, $\phi(\mathcal{F}, \delta, n)$ bounds the worst-case estimation error of $\mathbb{P}(|Y - f(X)| \leq t)$ using the empirical estimate $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{|Y_i - f(X_i)| \leq t\}$ over the whole function class $\mathcal{F}$ and for any value of t. Typically, $\phi(\mathcal{F}, \delta, n)$ will decrease with an increasing number of data points n and increase as the *complexity* of $\mathcal{F}$ gets larger.

Theorem 1. *Let $C^{1-\alpha}_{\hat{f},\hat{t}}(x)$ be the prediction interval output by `EffOrt` . If Assumption 1 and 2 are satisfied, the distribution of Y is atomless, n_c and n_ℓ are large enough so that $\frac{1-\alpha}{n_c} + \sqrt{\frac{\log(2/\delta)}{2n_c}} \leq r$ and $\phi(\mathcal{F}, \delta, n_\ell) \leq r$, then:*

1. $\mathbb{P}(Y \in C^{1-\alpha}_{\hat{f},\hat{t}}(X) | \mathcal{D}^{lrn}) \geq 1-\alpha$ a.s.

2. With probability greater that $1-2\delta$:

$$\lambda\left(C^{1-\alpha}_{\hat{f},\hat{t}}\right) \leq \lambda\left(C^{1-\alpha}_{f^*,t^*}\right) + 2L\Big(\frac{1}{n_c} + \sqrt{\frac{\log(2/\delta)}{2n_c}}\Big)^{\gamma} + 4L\phi(\mathcal{F}, \delta, n_\ell)^{\gamma} \tag{9}$$

To the best of our knowledge, Theorem 1 is one of the first to provide such a finite-sample upper bound on the excess-volume loss. It explicitly reveals the impact of the two split conformal steps of `EffOrt`. The two first error terms (involving n_c) match the bound of Proposition 1, and can be seen as the volume loss due to the calibration step. While the third term, with $\phi(\mathcal{F}, \delta, n_\ell)$, is the error due to the learning step. If we omit the dependence in δ, $\phi(\mathcal{F}, \delta, n_\ell)$ will typically be in the form of $\sqrt{\frac{\text{Compl}(\mathcal{F})}{n_\ell}}$, where $\text{Compl}(\mathcal{F})$ measures the complexity of $\mathcal{F}$. In most settings, we have $\text{Compl}(\mathcal{F}) \gg \log(1/\delta)$. Hence, the rate in Eq. (9) supports the important intuition that the learning step remains more important than the conformal step, at least in the sense that more data-points are needed to reach convergence. It is thus preferable to assign more points to the learning than the calibration.

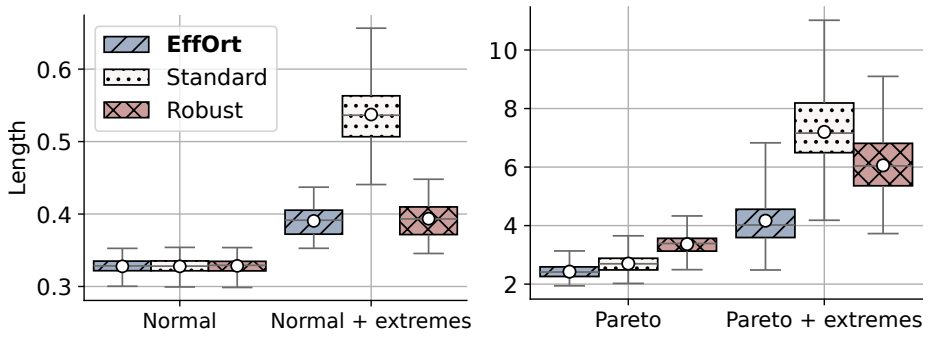


Figure 1: Boxplots of the 50 empirical expected lengths obtained by evaluating `EffOrt`. The white circle corresponds to the mean.

4 Experiments

We consider a linear regression model $Y = X^T\theta + \mathcal{E}$ where $\theta \sim \mathcal{U}(0,1)^{\otimes 3}$ is fixed, $X \sim \mathcal{N}(0, I_3)$ with I_3 the identity matrix of size 3×3, and $\mathcal{E}$ follows 4 different distributions: A standard normal, a mixture distribution $0.95 \cdot \mathcal{N}(0,1) + 0.05 \cdot \mathcal{N}(2,1)$, a Pareto distribution with shape and scale parameters equal to 2 and 1, and another mixture equals to $0.95 \cdot \text{Pareto}(2,1) + 0.05 \cdot \mathcal{N}(-20,1)$. In the two mixtures, the additional normal distributions allow simulating extreme values. For each scenario, we generate $n_{lrn} = n_{cal} = 1000$ pairs (X_i, Y_i), as well as $n_{test} = 1000$ test points to compute the empirical marginal coverage and the average size of the returned set. We repeat this procedure 50 times. During the learning step of `EffOrt`, the space of research $\mathcal{F}$ is restricted to the space of linear functions. For the split CP method, the regression function is either estimated using a linear regression or, in order to be fair in our comparisons, using a robust linear regression with Huber loss with parameter $\delta = 1.35$. For all methods, the score function is the absolute value of the residuals, i.e., $s(x,y) = |y - \widehat{f}(x)|$ and we set $\alpha = 0.1$. Overall, `EffOrt` produces more efficient marginally valid sets than those obtained with the split CP method, in all scenarios (see Fig. 1).

References

Angelopoulos, A. N. and Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer.

Pena-Ordieres, A., Luedtke, J. R., and Waachter, A. (2020). Solving chance-constrained problems via a smooth sample-based nonlinear approximation. *SIAM Journal on Optimization*, 30(3):2221–2250.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

On Strenghtenings for the Feature Selection Problem

Renaud CHICOISNE[1] & Pierre LATOUCHE [2,4] & Rui MA [3]

[1] *Université Clermont-Auvergne, CNRS, Mines de Saint-Etienne, Clermont-Auvergne-INP, LIMOS, France* renaud.chicoisne@uca.fr
[2] *Université Clermont-Auvergne, CNRS, laboratoire LMBP, France,* pierre.latouche@uca.fr
[3] *Université Clermont-Auvergne, CNRS, Mines de Saint-Etienne, Clermont-Auvergne-INP, LIMOS, France* rui.ma@doctorant.uca.fr
[4] *Institut Universitaire de France (IUF)*

Résumé. Le principe d'optimisation parcimonieuse, visant à résoudre le problème de sélection des caractéristiques, suscite un intérêt croissant dans les domaines des statistiques, de l'apprentissage automatique et de l'apprentissage profond. Dans cet article, nous nous concentrons sur l'optimisation parcimonieuse exacte, en intégrant la régularisation L_0 dans notre formulation. Nous introduisons une relaxation renforcée issue d'une convexification partielle dans l'espace original et proposons un algorithme général d'optimisation par plans coupants, adapté aux problèmes d'optimisation quadratique en nombres entiers mixtes. Notre méthode est évaluée sur des données simulées pour lesquelles des résultats sont présentés.

Mots-clés. Optimisation Parcimonieuse, Norme L_0, Optimisation Quadratique, Programmation en Nombres Entiers, Plans Coupants.

Abstract. The sparse optimization problem, which aims at addressing the feature selection problem, has garnered increasing attention in the fields of statistics, machine and deep learning. In this paper, we focus on exact sparse optimization by incorporating L_0 regularization into our formulation. We introduce a strong relaxation derived from partial convexifications in the original space and propose a general cutting-plane optimization algorithm tailored for quadratic mixed-integer optimization problems. Our algorithm is evaluated on simulated data, and the experimental results are presented.

Keywords. Feature Selection, $L_0 - norm$, Quadratic Optimization, Integer Programming, Cutting-plane Methods.

1 Introduction

The mean square error is one of the most widely used error function for measuring the difference between a regressor and a response vector. When facing high-dimensional data, obtaining sparse regressors is crucial as it helps to lower the dimension of the data, and gives steep insights on its structure, making it more interpretable.

In practice, to get an approximate sparse solution in reasonable time, the most common way to get a sparse regressor is to apply heuristic methods based on convex - and thus tractable - approximations. In the works of Latouche et al. (2016); Bouveyron et al. (2018, 2020), a simple relaxation along with an expectation-maximization algorithm were employed for inference. Besides, Lasso (Tibshirani, 1996) and Elastic net (Zou and Hastie, 2005) regularizations convexly penalize dense potential regressors. However, such methods can lead to biased regression results (Bertsimas et al., 2016). In contrast, the L_0 regularization - which directly controls the sparsity - is less used in the statistical community. More formally, given a set of m observations in an n-dimensional feature space represented by an $m \times n$ matrix X, a response vector $y \in \mathbb{R}^m$ and a parameter $\alpha > 0$, the L_0-regularized regression problem can be written as follows:

$$\min_{\beta} \frac{1}{2}||X\beta - y||_2^2 + \alpha||\beta||_0, \tag{1}$$

where the (nonconvex) L_0 term $||\cdot||_0$ "counts" the number of nonzero coefficients of the vector it takes as an argument. It is well-established that solving such a mixed-integer optimization problem (MIP) is NP-hard (Natarajan, 1995). In the optimization community, two main categories of algorithms are used to solve MIPs: exact and heuristic methods. Exact methods leverage the computational power of modern optimization solvers, solving MIPs through branch and bound algorithms. Additionally, when the matrix $Q = X^\top X$ is sparse, exact methods (Bertsimas et al., 2016; Hastie et al., 2015; Wainwright, 2009) tend to perform better. On the other hand, heuristic methods do not guarantee an optimal regressor but are significantly faster than exact methods. In practice, well-implemented packages (Friedman et al., 2023; Hazimeh and Mazumder, 2020) are readily available and can give reasonable regressors β, having, however, no guarantee of being optimal solutions to (1).

In this work, we focus on the exact solution of (1): Defining $a := -X^\top y$ and $b := \alpha \cdot \mathbf{1}$, it can be rewritten - up to a constant - as the following extended MIP:

$$\begin{aligned} \min_{\beta,z} \quad & \frac{1}{2}\beta^\top Q\beta + a^\top\beta + b^\top z \\ \text{s.t.} \quad & \beta_j(1 - z_j) = 0, \quad z_j \in \{0,1\}, \qquad \forall j \in [n], \end{aligned} \tag{2}$$

where $[n] = \{1, \ldots, n\}$. The complementarity constraints $\beta_j(1-z_j) = 0$ can be enforced using big-M constraints, which - unfortunately - are notoriously hard to deal with in MIPs, as they induce serious numerical issues and provide extremely weak relaxation bounds. A classic approach to strengthen big-M formulations consists in generating strong cutting planes and reoptimize until we obtain a satisfying bound. This approach is unfortunately not adequate here, as reoptimizing *nonlinear* MIPs such as (2) is not efficient in practice. On the other hand, *linear* MIPs are particularly easy to reoptimize: In this paper, we build up on two recent results of Wei et al. (2024): 1) their reformulation (MILO) of (2) that is even weaker than classical nonlinear big-M formulations but is a *linear* MIP, and 2) the computationally expensive convexification that led to MILO to generate strong cutting planes.

In Section 2 we generalize their convexification to lower dimensions to obtain a tradeoff between strength and computational efficiency and then use it to generate strong cutting planes for the linear MIP MILO, which can be reoptimized efficiently after being strengthened

by these new cuts. We present in Section 3 a comparison of the relaxation bounds obtained by our approach, the classical nonlinear big-M based MIP and MILO, on medium-sized synthetic instances. If this paper is accepted, extra results on large instances and on real data will be presented. In this work, we focus exclusively on the case where Q is positive definite. To secure this property, we incorporate an L_2 regularization term with a coefficient of 1 in the objective function.

2 Optimization

The main difficulty of (2) resides in the fact that the following set is nonconvex:

$$\mathcal{S} := \left\{(\beta, z, t) \in \mathbb{R}^n \times \{0,1\}^n \times \mathbb{R} : t \geqslant \beta^\top Q\beta, \quad \beta_j(1 - z_j) = 0, \forall j \in [n]\right\},$$

making (2) equivalent to $\min_{t,\beta,z}\{t/2 + a^\top\beta + b^\top z : (\beta, z, t) \in \mathcal{S}\}$. As optimizing a linear functional over a set is equivalent to do so over its convex hull, any information of said convex hull is of paramount importance: In Wei et al. (2024), they prove that $(\beta, z, t) \in \operatorname{conv} \mathcal{S}$ if and only if there exists a symmetric matrix $W \in \mathbb{S}^n$ such that:

$$\begin{aligned} &\begin{pmatrix} t & \beta^\top \\ \beta & W \end{pmatrix} \in \mathbb{S}_+^{n+1} \\ &(z, W) \in \operatorname{conv}\left\{\left(z^{\mathcal{N}}, Q^{\mathcal{N}}\right), \mathcal{N} \subset [n]\right\} \end{aligned} \tag{3}$$

where $\mathbb{S}_+^{n+1}$ is the cone of $(n+1) \times (n+1)$ symmetric semidefinite positive matrices, each vector $z^{\mathcal{N}} \in \{0,1\}^n$ is the indicator of $\mathcal{N}$ and each $n \times n$ matrix $Q^{\mathcal{N}}$ is the inverse of the submatrix of Q restricted to $\mathcal{N}$, padded with zeroes everywhere else.

Partial Convexification and Separation While the latter result gives a nice mathematical representation of conv $\mathcal{S}$, it is intractable in practice due to the presence of the convex hull (3). We now propose to partially convexify $\mathcal{S}$ in the following Theorem, for which the proof is left to the full length paper:

Theorem 1. *Given $\mathcal{J} \subset [n]$ of cardinality q, (β, z, t) belongs to the convex hull of*

$$\mathcal{S}_{\mathcal{J}} := \left\{(\beta, z, t) \in \mathbb{R}^n \times \{0,1\}^n \times \mathbb{R} : t \geqslant \beta^\top Q\beta, \quad \beta_j(1 - z_j) = 0, \forall j \in \mathcal{J}\right\}$$

if and only if, there exists $\mu \in \mathbb{R}$ and a matrix $W \in \mathbb{S}^q$ such that

$$\begin{aligned} &t \geqslant \beta^\top \begin{pmatrix} T^\top S^{-1} T & T^\top \\ T & S \end{pmatrix} \beta + \mu, \qquad \begin{pmatrix} \mu & \beta(\mathcal{J})^\top \\ \beta(\mathcal{J}) & W \end{pmatrix} \in \mathbb{S}_+^{q+1} \\ &(z(\mathcal{J}), W) \in \mathit{conv}\left\{\left(z^{\mathcal{N}}, \left(P - T^\top S^{-1} T\right)^{\mathcal{N}}\right), \mathcal{N} \subset \mathcal{J}\right\} \end{aligned} \tag{4}$$

where $z(\mathcal{J})$ and $\beta(\mathcal{J})$ are respectively the vectors z and β restricted to their components in $\mathcal{J}$ and Q is block-decomposed as follows, where wlog the first rows and columns correspond to the indices in $\mathcal{J}$:

$$Q = \begin{pmatrix} P & T^\top \\ T & S \end{pmatrix}$$

Theorem 1 allows us to partially strengthen any formulation for our main problem (2). For reasonable sizes of $|\mathcal{J}| = q$, it is tractable to partially convexify $\mathcal{S}$, as (4) can be completely defined by enumerating all 2^q extreme points, as opposed to the 2^n in (3). By using duality results, we now present a way to separate an incumbent from conv $\mathcal{S}_\mathcal{J}$:

Corollary 1. *$(\beta, z, t) \in$ conv $\mathcal{S}_\mathcal{J}$ if and only if*

$$t \geqslant \beta^\top \begin{pmatrix} T^\top S^{-1} T & T^\top \\ T & S \end{pmatrix} \beta + \max_{\pi,\lambda,\lambda_0} \quad \pi^\top \beta(\mathcal{J}) + \lambda^\top z(\mathcal{J}) + \lambda_0$$
$$s.t. \quad \frac{1}{4}\pi^\top \left(P - T^\top S^{-1} T\right)^{\mathcal{N}} \pi + \lambda_0 + \sum_{j \in \mathcal{N}} \lambda_j \leqslant 0, \quad \forall \mathcal{N} \subset \mathcal{J}$$

Given a (fractional) incumbent $(\bar{\beta}, \bar{z}, \bar{t})$, we can generate a cut separating it from conv $\mathcal{S}_\mathcal{J}$ by solving the (moderately large) maximization problem above, which can be done in polynomial time via interior point methods (Potra and Wright, 2000). Letting $(\pi^*, \lambda^*, \lambda_0^*)$ be its optimal solution, the following cut is valid for (2):

$$t \geqslant \beta^\top \begin{pmatrix} T^\top S^{-1} T & T^\top \\ T & S \end{pmatrix} \beta + \beta(\mathcal{J})^\top \pi^* + z(\mathcal{J})^\top \lambda^* + \lambda_0^*,$$

In practice, the selection of $\mathcal{J}$ is crucial. We discuss this in more detail in Section 3. As mentioned earlier, while these cuts are valid, adding them iteratively and reoptimize many times cannot be done efficiently in this *nonlinear* setting. We now show that we can translate these partial enumeration cuts for the *linear* MILO formulation from Wei et al. (2024).

Linear Reformulation and Equivalent Cuts Wei et al. (2024) recast the strong, but exponential-size representation (3) into the following weak, but polynomial-size linear MIP that is also devoid of variables β:

$$(MILO) \qquad \begin{aligned}
\min_{z,W,\eta} \quad & \eta \\
\text{s.t.} \quad & \eta \geqslant -\frac{1}{2} a^\top W a + b^\top z \\
& \sum_{k=1}^{n} Q_{ik} W_{ki} = z_i, && \forall i \in [n] && (5) \\
& -M(1 - z_i) \leqslant \sum_{k=1}^{n} Q_{ik} W_{kj} \leqslant M(1 - z_i), && \forall i \neq j && (6) \\
& -\lambda_{\max}(Q^{-1}) z_i \leqslant W_{ij} \leqslant \lambda_{\max}(Q^{-1}) z_i, && \forall (i,j) \in [n]^2 && (7) \\
& -\lambda_{\max}(Q^{-1}) z_j \leqslant W_{ij} \leqslant \lambda_{\max}(Q^{-1}) z_j, && \forall (i,j) \in [n]^2 && (8) \\
& z \in \{0,1\}^n, \quad W \in \mathbb{S}^n && && (9)
\end{aligned}$$

Where $M := \lambda_{\max}(Q^{-1}) \cdot \max_{i \in [n]} ||Q_{[n],i}||_2$. Importantly, it is shown that the constraints (5)-(9) precisely define the exponentially defined set (3), and that we can retrieve the optimal regressor with $\beta^* = -W^* a$ from an optimal solution (z^*, W^*, η^*) of MILO. We now state our main result, which is the direct translation of the *nonlinear* separation problem from Corollary 1, now in terms of the MILO formulation:

Corollary 2. *For any set $\mathcal{J} \subset [n]$, the following inequality is valid for MILO:*

$$\eta \geqslant -\frac{1}{2}\bar{a}^\top S^{-1}\bar{a} + \max_{\lambda,\lambda_0} \lambda^\top z(\mathcal{J}) + \lambda_0$$
$$s.t.\ \lambda_0 + \sum_{j\in\mathcal{J}, z_j^{\mathcal{N}}=1} \lambda_j + \frac{1}{2}(\hat{a} - T^\top S^{-1}\bar{a})^\top \hat{W}^{\mathcal{N}}(\hat{a} - T^\top S^{-1}\bar{a}) \leqslant 0,\ \forall \mathcal{N} \subset \mathcal{J},$$

where $\hat{a} := a(\mathcal{J})$, $\bar{a} := a(\mathcal{J} \setminus [n])$, $\hat{W} := P - T^\top S^{-1} T$ and P, T, S the blocks of Q wrt $\mathcal{J}$.

Notice that the maximization problem above is *linear* and generates the following cuts:

$$\eta \geqslant -\frac{1}{2}\bar{a}^\top S^{-1}\bar{a} + z(\mathcal{J})^\top \lambda^* + \lambda_0^*.$$

3 Experiments on simulated data

We now compare the relaxation bounds obtained with the classic nonlinear big-M formulation and our MILO-based, cut-enhanced models: i.e. we compare the strength of the relaxations obtained by each model when the integrality of the $z \in \{0,1\}^n$ is relaxed to $z \in [0,1]^n$. We solve our models using CPLEX 12.10 (Cplex, 2009) with default settings on a desktop computer with Intel Core i5-10505 and 16 GB of RAM.

Simulated Data We simulated datasets (X, y) with two generators $\mathbb{G}1$ and $\mathbb{G}2$. For each experiment size and generator, we sample 10 datasets with identical settings and record the expected value ($\mathbb{E}$) and standard deviation (σ) of the improvement achieved (in %) by our methods against the classical nonlinear big-M model, as well as the running time. All the data sets are generated by sampling $\beta \sim \mathcal{N}(0_n, I_n/\tau_1)$ with a noise $\varepsilon \sim \mathcal{N}(0_m, I_m/\tau_2)$ and by computing $y = X\beta + \varepsilon$. The matrix X itself is generated from a Gaussian distribution with zero mean and a covariance matrix R. For $\mathbb{G}_1$, R is set to $R = \text{diag}(R^1, R^2)$, a 2-blocks diagonal matrix, where R^l is such that $R^l_{ii} = 1$ and $R^l_{ij} = \rho$ for $i, j \in [n/2]$ and $i \neq j$. For $\mathbb{G}_2$, a matrix $\tilde{R}$ is first built with each element sampled randomly and independently. Second, each element is set to 10 with probability 1/2 and to 0.01 otherwise. R is finally built by symmetrizing $\tilde{R}$ and by setting each diagonal element to $R_{ii} = \sum_{j=1}^n R_{ij} + 1$. For both $\mathbb{G}_1$ and $\mathbb{G}_2$, we set $\tau_1 = \tau_2 = 1$ and $\rho = 0.75$.

Set $\mathcal{J}$ Selection Given a fractional incumbent $\bar{z}$, we jointly used three heuristic choices for selecting the set $\mathcal{J}$ on which we will partially enumerate: 1) wrt the correlation of Q in (2), 2) wrt the "fractionality" of each $\bar{z}_j$, or 3) a random selection of indices. We find out that in practice, our cuts are not always improving the bound. Let T be the maximum number of non-contributing cuts that we apply before updating $\mathcal{J}$. We consider two macro-strategies for selecting $\mathcal{J}$: The first approach uses a fixed value of q, small enough so that 2^q is enumerable, for which we take $q \in \{2, ..., 6\}$ and the selection process for $\mathcal{J}$ terminates when T reaches 20. The second approach allows for a dynamic adjustment of q. In this

case, q is incremented by 1, until it reaches $q = 6$. At that point the cutting-plane algorithm terminates. We tested this second strategy with $T \in \{1, 10, 20\}$.

	(n, m)		$q = 2$	$q = 3$	$q = 4$	$q = 5$	$q = 6$	$T = 5$	$T = 10$	$T = 20$
$\mathbb{G}1$	(10,50)	$\mathbb{E}$	0.03	2.46	4.27	6.84	8.21	7.43	7.85	8.05
		σ	4.04	5.32	6.88	7.70	8.49	8.95	8.69	8.32
	(20,100)	$\mathbb{E}$	-2.51	-1.47	-0.6	0.6	1.2	0.73	0.79	0.83
		σ	1.29	1.41	1.47	1.71	2.23	2.49	2.41	2.54
$\mathbb{G}2$	(10,50)	$\mathbb{E}$	10.01	15.23	15.36	17.58	22.57	14.55	22.24	24.03
		σ	27.59	28.32	28.08	28.72	30.44	27.28	31.48	32.22
	(20,100)	$\mathbb{E}$	139.34	140.36	144.68	148.93	153.29	144.07	147.14	155.36
		σ	233.81	236.42	237.41	247.67	253.58	234.47	243.30	255.04

Table 1: Relaxation bound improvement (%) from adding cuts.

The values (%) presented in Table 1 are calculated as $100 \times \frac{\text{(MILO\&CUTS bound)}-\text{(BigM bound)}}{\text{(BigM bound)}}$. Unsurprisingly, we observe a general improvement of the relaxation bound as q increases.

	(n, m)		$q = 2$	$q = 3$	$q = 4$	$q = 5$	$q = 6$	$T = 5$	$T = 10$	$T = 20$
$\mathbb{G}1$	(10,50)	$\mathbb{E}$	3.97	4.43	4.17	4.92	4.76	5.54	9.36	18.09
		σ	0.95	1.12	0.91	1.85	1.68	0.31	0.83	3.65
	(20,100)	$\mathbb{E}$	54.30	53.06	51.49	49.21	51.38	59.77	94.1	157.19
		σ	21.52	15.97	15.62	13.94	18.39	15.09	22.44	46.92
$\mathbb{G}2$	(10,50)	$\mathbb{E}$	4.54	4.54	4.73	4.20	3.02	6.48	14.49	11.72
		σ	2.26	2.44	2.92	2.17	1.93	4.06	7.07	8.90
	(20,100)	$\mathbb{E}$	39.8	37.58	38.64	41.94	43.06	45.35	63.6	96.65
		σ	12.62	16.78	14.66	12.6	21.34	17.31	22.12	30.81

Table 2: Running times (in seconds) for each data setting.

As shown in Table 2, our cutting-plane method generally terminates in a reasonable amount of time. In particular, when compared to the time required to solve the the nonlinear big-M model, which takes an average of around 500 seconds, our method demonstrates a significant reduction in computation time.

4 Conclusions

In this paper, we present a partial convexification method for solving quadratic mixed-integer optimization problems, and used it to provide a new optimization scheme of the mean square error function under L_0 penalization . We also introduced a cutting-plane algorithm based on this method, which strengthens the linear reformulation MILO proposed by Wei et al. (2024) and yields favorable results compared to the conventional big-M formulation within a reasonable time. The work presented in this paper is part of an ongoing work. In the next phase, we will conduct experiments with larger values of q and real-life data.

References

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.

Charles Bouveyron, Pierre Latouche, and Pierre-Alexandre Mattei. Bayesian variable selection for globally sparse probabilistic pca. *Electronic Journal of Statistics*, 2018.

Charles Bouveyron, Pierre Latouche, and Pierre-Alexandre Mattei. Exact dimensionality selection for bayesian pca. *Scandinavian Journal of Statistics*, 47(1):196–211, 2020.

IBM ILOG Cplex. V12. 1: User's manual for CPLEX. *International Business Machines Corporation*, 46(53):157, 2009.

Jerome Friedman, Trevor Hastie, Rob Tibshirani, Balasubramanian Narasimhan, Kenneth Tay, Noah Simon, Junyang Qian, and James Yang. glmnet: Lasso and elastic-net regularized generalized linear models. *Astrophysics Source Code Library*, pages ascl–2308, 2023.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143(143):8, 2015.

Hussein Hazimeh and Rahul Mazumder. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5):1517–1537, 2020.

Pierre Latouche, Pierre-Alexandre Mattei, Charles Bouveyron, and Julien Chiquet. Combining a relaxed EM algorithm with occam's razor for bayesian variable selection in high-dimensional regression. *Journal of Multivariate Analysis*, 146:177–190, 2016.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

Florian A Potra and Stephen J Wright. Interior-point methods. *Journal of computational and applied mathematics*, 124(1-2):281–302, 2000.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Mark Wainwright. *Photosensitisers in biomedicine*. John Wiley & Sons, 2009.

Linchuan Wei, Alper Atamtürk, Andrés Gómez, and Simge Küçükyavuz. On the convex hull of convex quadratic optimization problems with indicators. *Mathematical Programming*, 204(1):703–737, 2024.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

Données de survie, données censurées

Survival Models: Proper Scoring Rule and Stochastic Optimization with Competing Risks

Julie Alberge[1], Vincent Maladière[2], Olivier Grisel[2], Judith Abécassis[1], Gaël Varoquaux[1]

[1] *SODA Team, Inria Saclay, Palaiseau, France, julie.alberge@inria.fr*
[2] *:probabl., Paris, France*

Résumé. Lorsqu'on étudie des données censurées, c'est à dire où certains labels manquent en raison d'une période d'observation limitée, l'analyse de survie porte sur la prédiction du temps de l'événement d'intérêt. La présence de plusieurs événements d'intérêt conduit à une variante de classification: prédire l'événement le plus probable, domaine moins exploré nommé risques compétitifs. Habituellement, les modèles de risques compétitifs couplent l'architecture avec la métrique utilisée, entraînant un coût computationnel trop important pour être déployés sur des grandes bases de données, de plus en plus courantes en santé publique.

Ainsi, nous proposons une nouvelle métrique strictement "proper", ajustée pour la censure et séparable, permettant une optimisation sur un sous-ensemble des données, chaque observation étant évaluée indépendamment. Notre métrique estime les probabilités des résultats et permet une optimisation stochastique pour les risques compétitifs, ce qui nous permet d'utiliser des arbres de gradient boosting efficacement. SurvivalBoost obtient de meilleurs résultats sur différentes métriques comparé à 12 modèles à l'état de l'art sur plusieurs métriques et sur plusieurs jeux de données réelles, en analyse de survie et risques compétitifs, avec une excellente calibration, la capacité de prédire à n'importe quel horizon, ainsi que des couts de calcul plus avantageux que les méthodes existantes.

Mots-clés. Survie, Risques Compétitifs, "Proper" métrique, Gradient-Boosting

Abstract. When dealing with right-censored data, where some outcomes are missing due to a limited observation period, survival analysis —known as *time-to-event analysis*— focuses on predicting the time until an event of interest occurs. Multiple classes of outcomes lead to a classification variant: predicting the most likely event, a less explored area known as *competing risks*. Classic competing risks models couple architecture and loss, limiting scalability.

To address these issues, we design a strictly proper censoring-adjusted separable scoring rule, allowing optimization on a subset of the data as each observation is evaluated independently. The loss estimates outcome probabilities and enables stochastic optimization for competing risks, which we use for efficient gradient boosting trees. **SurvivalBoost** not only outperforms 12 state-of-the-art models across several metrics on 4 real-life datasets, both in competing risks and survival settings, but also provides great calibration, the ability to predict across any time horizon, and computation times faster than existing methods.

Keywords. Survival Analysis, Competing Risks, Proper Scoring Rule, Gradient-Boosting

1 Introduction and Related Work

We all die at some point. Some applications call for predicting not *if* but *when* an event of interest is likely to occur. In such a setting of *time-to-event regression*, samples often have unobserved outcomes, *e.g.* individuals that have not been followed long enough for the event of interest to occur. Limiting the analysis to fully observed samples creates a censoring bias. To address this, *survival analysis* models use dedicated corrections for censorship.
Competing risks analysis generalizes survival analysis to multiple events, determining which will happen first. As the risks of the various events are seldom independent–for example, cancer and cardiovascular disease share inflammation or age risk factors, competing risks cannot be solved by running a survival model for each event.

Competing risks Competing risks, involving multiple possible outcomes, require new methods that can naturally adapt to the simpler survival analysis setting. Derived from the Kaplan and Meier [1958] estimator, the Nelson-Aalen et al. [2008] estimator is an unbiased marginal model for competing risks. The linear Fine and Gray [1999] estimator, inspired by the Cox [1972] estimator in survival analysis, is the most popular model in clinical research. Recently, machine learning models have been adapted to competing risks settings, including neural networks approaches (e.g. DeepHit [Lee et al., 2018]).

2 Problem formulation

Notations We write oracle quantities as a^*, estimates as $\hat{a}$, vectors as $\mathbf{a}$, random variables in upper case, A, observations in lower cases a, and distributions in calligraphic style $\mathcal{A}$.
We consider $K \in \mathbb{N}^*$ competing events. For $k \in [\![1, K]\!]$, we denote $T_k^* \in \mathbb{R}_+$ the event time of the event k, depending on the covariates $\mathbf{X} \sim \mathcal{X}$. We also denote $T^* \in \mathbb{R}_+$, the first event of interest that occurs, $T^* = \min_{k \in [\![1,K]\!]}(T_k^*)$. We observe $(\mathbf{X}, T, \Delta) \sim \mathcal{D}$, with $T = \min(T^*, C)$ where $C \in \mathbb{R}_+$ is the censoring time, which can depend on $\mathbf{X}$, and $\Delta \in [\![0, K]\!], \Delta = \arg\min_{k \in [\![0,K]\!]}(T_k^*)$, where 0 denotes a censored observation. However, we are primarily interested in the distribution of the uncensored data, $(\mathbf{X}, T^*, \Delta) \sim \mathcal{D}^*$, particularly the joint distribution of $T^*, \Delta | \mathbf{X} = \mathbf{x}$.
Given a data set of n individuals, we denote each individual i by its associated covariates $\mathbf{x}_i$. The outcome is represented by (t_i, δ_i), where t_i is the observed time, and $\delta_i \in [\![0, K]\!]$ is the event indicator. $\delta_i = k$ indicates that the event of interest k was observed at time t_i, while $\delta_i = 0$ indicates that the observation was censored at time t_i. This paper aims to predict all cause-specific Cumulative Incidence Functions (CIFs) at any time horizon ζ.

Definition 2.1 (*Quantities of interest*)**.**

CIF (Cumulative Incidence Function): $S^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* > \zeta|\mathbf{X} = \mathbf{x})$

CIF (Cumulative Incidence Function): $F^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* \leq \zeta|\mathbf{X} = \mathbf{x}) = 1 - S^*(\zeta|\mathbf{X} = \mathbf{x})$

CIF of the k^{th} event: $F_k^*(\zeta|\mathbf{x}) = \mathbb{P}(T^* \leq \zeta \cap \Delta = k|\mathbf{X} = \mathbf{x})$

Censoring Function: $G^*(\zeta|\mathbf{x}) = \mathbb{P}(C > \zeta|\mathbf{X} = \mathbf{x})$

Assumption 2.1. We make the classic assumption in survival analysis: $T^* \perp\!\!\!\perp C \,|\, \mathbf{X}$

2.1 A strictly proper scoring rule for competing risks

Proper Scoring Rule Scoring rules are cost functions of observations and a candidate probability distribution. A scoring rule ℓ evaluates a distribution $\mathcal{P}$ on an observation Y, producing a corresponding score $\ell(\mathcal{P}, Y)$. The higher the score, the better the model fits the observation. For a proper scoring rule, the score reflects the model's ability to predict the oracle distribution.

Definition 2.2 (*Proper Scoring Rule*)**.** A scoring rule ℓ is considered proper if

$$\forall \mathcal{P}, \mathcal{Q}, \text{distributions} \quad \mathbb{E}_{Y\sim\mathcal{Q}}[\ell(\mathcal{P}, Y)] \leq \mathbb{E}_{Y\sim\mathcal{Q}}[\ell(\mathcal{Q}, Y)].$$

If the equality holds if and only if $\mathcal{P} = \mathcal{Q}$, in which case the scoring rule is *strictly proper*.

Proper scoring rule for the Global CIF We denote L_ζ a scoring rule for the global CIF at time ζ.

Definition 2.3 (*PSR for competing risks settings*)**.** In competing risks settings, where censoring is present, a scoring rule L_ζ for the CIF at time ζ for an observation $(\mathbf{X}, T, \Delta)$ is proper if and only if: $\forall \zeta, (\mathbf{X}, T, \Delta) \sim \mathcal{D}, \forall (\hat{F}_1, ..., \hat{F}_K, \hat{S})$,

$$\begin{aligned} & \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}[L_\zeta(\ (\hat{F}_1(\zeta|\mathbf{x}), ..., \hat{F}_K(\zeta|\mathbf{x}), \hat{S}(\zeta|\mathbf{x}))\ , (T, \Delta))] \\ \leq\ & \mathbb{E}_{T^*,C,\Delta|\mathbf{X}=\mathbf{x}}[L_\zeta(\ (F_1^*(\zeta|\mathbf{x}), ..., F_K^*(\zeta|\mathbf{x}), S^*(\zeta|\mathbf{x}))\ , (T, \Delta))] \end{aligned} \tag{1}$$

Estimated distributions

Oracle distributions

When equality is achieved *only* for the oracle distributions, the scoring rule is *strictly proper*.

Definition 2.4 (Competitive Weights Negative LogLoss)**.** We introduce the multiclass negative log-likelihood, re-weighted with the censoring distribution. $\forall \zeta, (\mathbf{x}, t, \delta) \sim \mathcal{D}$

$$\mathrm{L}_\zeta((\hat{F}_1(\zeta|\mathbf{x}), ..., \hat{F}_K(\zeta|\mathbf{x}), \hat{S}(\zeta|\mathbf{x})), (t, \delta)) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\frac{\mathbb{1}_{t_i\leq\zeta,\delta_i=k}\ \log\left(\hat{F}_k(\zeta|\mathbf{x}_i)\right)}{G^*(t_i|\mathbf{x}_i)}\right) + \frac{\mathbb{1}_{t_i>\zeta}\ \log\left(\hat{S}(\zeta|\mathbf{x}_i)\right)}{G^*(\zeta|\mathbf{x}_i)} \tag{2}$$

Probability of remaining censor-free at t_i

Probability of remaining censor-free at ζ (1 - probability of censoring)

Theorem 1 (Properness of the scoring rule)**.** *Under the assumption that the weights are appropriately chosen, $L_\zeta : \mathbb{R}^{K+1} \times \mathcal{D} \to \mathbb{R}$ is a strictly proper scoring rule for the global CIF on a fixed time horizon $\zeta \in \mathbb{R}_+$.*

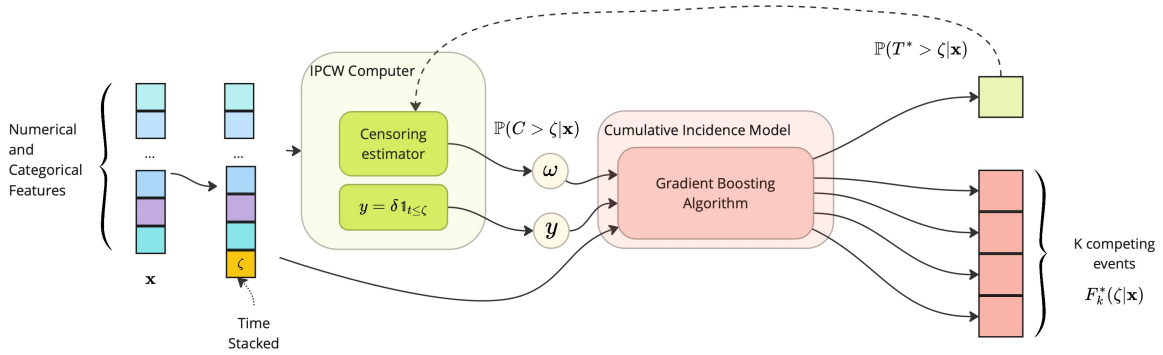


Figure 1: **SurvivalBoost Algorithm with its Feedback Loop.**

3 SurvivalBoost: Gradient Boositng competing risks

While eq.2 can be used as a loss in any multiclass machine learning algorithm, we choose Gradient Boosting Trees due to their strong performance on tabular data. We introduce an algorithm called SurvivalBoost, which predicts all CIFs for each competing event as well as the global survival function. Our algorithm utilizes two classifiers (here, gradient-boosted trees), one for censoring, trained on binary censored/non-censored labels (i.e., for time ζ, $\mathbb{P}(C > \zeta|\mathbf{X} = \mathbf{x})$), and one for multiple events.
Figure 1 illustrates an iteration: we compute the weights w_i and targets y_i based on the sampled times for each individual (eq. 2). Specifically, for censored samples, the corresponding weight is set to 0. A target $y_i \in [\![1, K]\!]$ indicates that the event of interest occurred before ζ and when $y_i = 0$, the individual has survived without experiencing any event.

4 Experiments

We present the datasets, evaluation metrics and the baselines on the competing risks and survival analysis settings. Then, we present the results obtained for both settings.

Evaluating the predicted probability We extend the Brier Score in competing risks proposed by Graf et al. [1999] with a formal proof of the properness of the loss.

Competing Risks: SEER Dataset This dataset tracks 470,000 breast cancer patients for up to ten years, with mortality due to various diseases as the outcomes. The censoring rate is approximately 63%.

Survival Analysis: KKBOX Dataset The Churn Prediction Challenge 2017 hosted on Kaggle, which features administrative censoring and 2.5M data points. We trained the models over 100k, 1M, and 2M data points to assess scalability.

Competing risks: Baselines We compare our approach with 7 other competing risks models from simpler models: Aalen et al. [2008]'s and the Fine and Gray [1999] linear model. We benchmark against tree-based approach - Random Survival Forests (RSF)

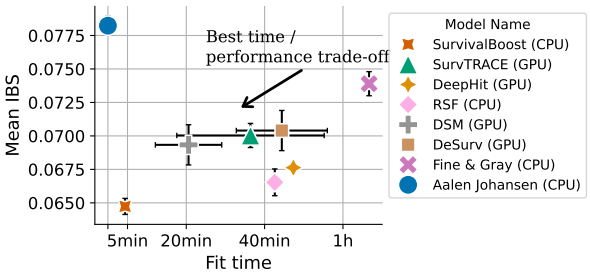


Figure 2: **Prediction performance / training time trade-off for competing risks on SEER dataset**. Average IBS versus fitting time for each model.

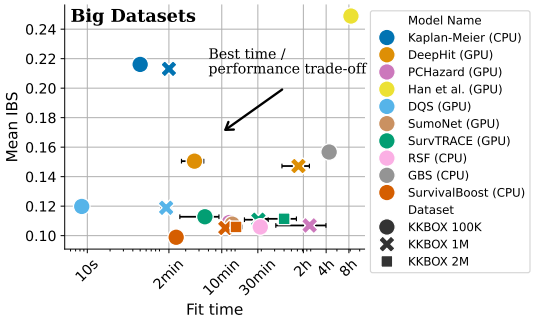


Figure 3: **Prediction performance / training time trade-off in survival analysis** IBS (Integrated Brier score) function of fit time for each model on KKBOX.

[Ishwaran et al., 2008] and several neural network-based models with DeepHit [Lee et al., 2018], Deep Survival Machines [Nagpal et al., 2021] DeSurv [Danks and Yau, 2022], and SurvTRACE [Wang and Sun, 2022].

Survival Analysis: Baselines We benchmark our method against the most performant SOTA survival models. This includes PCHazard [Kvamme and Borgan, 2019], SumoNet [Rindt et al., 2022], and DQS [Yanagisawa, 2023]. We also evaluate survival games [Han et al., 2021], and Gradient Boosting Trees methods with GBS - from Scikit-survival.

Results on SEER Dataset Figure 2 compares the models using the Integrated Brier Score (with Kaplan-Meier weights from Graf et al. [1999] due to the absence of an oracle). SurvivalBoost achieves both the best score and the shortest training time.

Results on KKBOX Figure 3 shows the trade-off between training time and performance in terms of IBS, where SurvivalBoost excels, being the top model in statistical performance and one of the fastest on KKBOX.

Code reproducibility The code is available on GitHub as a library: hazardous.

Conclusion For competing risks, which generalizes survival analysis to classify the type of outcome, we first propose and prove a strictly proper scoring rule. This reweighted log loss can easily be used in machine learning models: it is separable by observation, making it suitable for stochastic solvers, it does not require time derivatives and it can be applied to non-differentiable models. We integrate it into gradient-boosting trees, resulting in an algorithm called SurvivalBoost. By using time as a feature and incorporating a feedback loop to better estimate censoring probabilities, SurvivalBoost outperforms state-of-the-art methods real-life datasets, for both competing risks and survival analysis. It also trains faster on large datasets. As a loss function, it allows survival analysis or competing risks modeling to be easily extended to a wide range of models replacing clinical standards that do not scale.

References

Odd O. Aalen, Ornulf Borgan, and Haakon K. Gjessing. *Survival and Event History Analysis*. Statistics for Biology and Health. Springer New York, 2008.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Dominic Danks and Christopher Yau. Derivative-Based Neural Modelling of Cumulative Distribution Functions for Survival Analysis. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, 2022.

Jason P. Fine and Robert J. Gray. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18 (17-18):2529–2545, 1999.

Xintian Han, Mark Goldstein, Aahlad Puli, Thomas Wies, Adler Perotte, and Rajesh Ranganath. Inverse-weighted survival games. *Advances in neural information processing systems*, 34:2160–2172, 2021.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3), 2008.

E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Haavard Kvamme and ornulf Borgan. Continuous and Discrete-Time Survival Prediction with Neural Networks, 2019. arXiv:1910.06724.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Chirag Nagpal, Xinyu Rachel Li, and Artur Dubrawski. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks, 2021.

David Rindt, Robert Hu, David Steinsaltz, and Dino Sejdinovic. Survival Regression with Proper Scoring Rules and Monotonic Neural Networks, 2022. arXiv:2103.14755.

Zifeng Wang and Jimeng Sun. SurvTRACE: Transformers for Survival Analysis with Competing Events. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–9, 2022.

Hiroki Yanagisawa. Proper scoring rules for survival analysis, 2023.

A general representation of conditional hazard functions in survival analysis

Daphné Aurouet Valentin Patilea

Univ Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France
{*daphne.aurouet, valentin.patilea*}*@ensai.fr*

Résumé. Dans cet article, nous dérivons une représentation simple de la fonction de hasard conditionnel en utilisant une densité jointe et une espérance conditionnelle. Il est démontré que ces représentations simples de la fonction de hasard conditionnel sont disponibles pour plusieurs types de censure aléatoire, de troncature et en présence d'individus immunisés à l'évènement, ce qui ouvre la voie à des approches d'estimation non paramétriques dans de nombreux modèles prédictifs.

Mots-clés. Modèle de cure, Current status, Censure à gauche ou droite, Troncation.

Abstract. In this paper we derive a simple representation of the conditional hazard function using a joint density and a conditional expectation. It is shown that such simple representations of the conditional hazard function are available under several common types of random censoring, truncation, and in the presence of cured individuals, opening the door to nonparametric estimation approaches in many time-to-event predictive models.

Keywords. Cure models, Current status, Left or right censoring, Truncation.

1 Introduction

Time-to-event analysis, also called event history, is a long-standing topic in statistics that deals with the time to a particular event or outcome. Application fields include, but are not limited to, biostatistics, engineering, labour economics, finance, insurance, and demography. In the most common setting, the variable of interest is a non-negative random variable T representing the time until the event of interest occurs. The aim is to estimate the distribution of this variable, or its conditional distribution if predictors are also observed.

The main challenge in time-to-event analysis comes from the fact that the variable T is not always observed. For a fraction or all of the observed individuals, the data only indicate that T is less than or greater than some observed value, or that T is between two observed values. The existence of some individuals may even be unknown, or the value of their T is infinite. We shall call all reasons that prevent the exact observation of T: *incompleteness mechanisms*. Thus they include censoring (left, right, interval), truncation (left, right), and long-time survivors (cured individuals). See the textbooks and reviews Klein and Moeschberger (2003), Kleinbaum and Klein (2012), Klein et al. (2014), Amico

and Van Keilegom (2018) for extensive lists of different types of common incompleteness mechanisms in time-to-event analysis.

The modeling of the distribution of the time-to-event T is often done using the hazard function, especially when predictors are available, in which case the interest is in the conditional hazard function given the predictors (also called covariates). The classical approaches to modeling the hazard function are parametric and semiparametric as in the case of the famous Cox's proportional hazard model. However, model-free (also called nonparametric) approaches are preferable when the underlying assumptions of the (semi)parametric models do not meet consensus.

A key message of this paper is that conditional hazard functions allow for a user-friendly representation as a ratio between the joint density and a conditional expectation. This representation is known in the case of right-censoring, see for example Efromovich (2018), Efromovich (2024). We show that a similar representation can be derived in many situations where an incompleteness mechanism prevents the observation of T for all individuals. Our examples include models with left or right random censoring, with or without random truncation, with or without a cure fraction, and also some types of modified current status models. If such ratio-type representations of conditional hazard can be established, this opens the door to the use of appropriate nonparametric estimators of the numerator and denominator. In particular those adapted to massive data sets, possibly arriving in streams. See Aurouet and Patilea (2025) for a recursive kernel estimation approaches.

2 Methodology

In our framework, the data include a positive, continuous random variable T, representing the survival time of interest, and a vector $\mathbf{X} = (\mathbf{X}_c, \mathbf{X}_d)$ of covariates. The sub-vector $\mathbf{X}_c \in \mathcal{X}_c \subset \mathbb{R}^{d_c}$ contains $d_c \geq 1$ continuous predictors, and $\mathbf{X}_d \in \mathcal{X}_d$ gathers the $d_d \geq 0$ discrete predictors. Let $\mathcal{X}$ be the support of $\mathbf{X}$.

2.1 Random right censoring

In the case of right censoring, instead of T, the observed variables are the duration $Y = T \wedge C$ together with the indicator $\delta = \mathbb{1}\{T \leq C\}$, where $C \geq 0$ is the censoring variable. The data are represented by the triplets $(Y_i, \delta_i, \mathbf{X}_i)$ for $i = 1, \ldots, n$, independent copies of $(Y, \delta, \mathbf{X})$. We consider the following common identification condition.

Assumption 1. *T and C are conditionally independent given $\boldsymbol{X}$ (notation : $T \perp C \mid \boldsymbol{X}$).*

Let $F_{T|\mathbf{X}}(t \mid \boldsymbol{x}) = \mathbb{P}(T \leq t \mid \mathbf{X} = \boldsymbol{x})$ and $F_{C|\mathbf{X}}(t \mid \boldsymbol{x}) = \mathbb{P}(C \leq t \mid \mathbf{X} = \boldsymbol{x})$, for $t \geq 0$ and $\boldsymbol{x} \in \mathcal{X}$, denote the conditional distribution functions of T and C given $\mathbf{X} = \boldsymbol{x}$, respectively. Let $S_{T|\mathbf{X}}(t \mid \boldsymbol{x}) = 1 - F_{T|\mathbf{X}}(t \mid \boldsymbol{x})$ and $S_{C|\mathbf{X}}(t \mid \boldsymbol{x}) = 1 - F_{C|\mathbf{X}}(t \mid \boldsymbol{x})$, be the corresponding survival functions. Moreover, for each $\boldsymbol{x} \in \mathcal{X}$, we also consider the conditional sub-distributions functions of Y given $\mathbf{X} = \boldsymbol{x}$, that are defined as

$$H_k(t \mid \boldsymbol{x}) = \mathbb{P}(Y \leq t, \delta = k \mid \mathbf{X} = \boldsymbol{x}), \qquad k \in \{0, 1\}, \tag{1}$$

and the conditional distribution function as

$$H(t \mid \boldsymbol{x}) = \mathbb{P}(Y \leq t \mid \mathbf{X} = \boldsymbol{x}) = H_0(t \mid \boldsymbol{x}) + H_1(t \mid \boldsymbol{x}). \tag{2}$$

By Assumption 1, we have $1 - H(t \mid \boldsymbol{x}) = S_{T|\mathbf{X}}(t \mid \boldsymbol{x})S_{C|\mathbf{X}}(t \mid \boldsymbol{x})$. It is well known that the conditional hazard and survival functions given $\mathbf{X} = \boldsymbol{x}$ can be expressed as

$$\lambda_{T|\mathbf{X}}(t \mid \boldsymbol{x}) := \frac{f_{T|\mathbf{X}}(t \mid \boldsymbol{x})}{S_{T|\mathbf{X}}(t \mid \boldsymbol{x})} = \frac{f_{Y,\delta|\mathbf{X}}(t, 1 \mid \boldsymbol{x})}{1 - H(t- \mid \boldsymbol{x})} = \frac{f_{Y,\delta,\mathbf{X}}(t, 1, \boldsymbol{x})}{R(t, \boldsymbol{x})}, \tag{3}$$

where $f_{Y,\delta|\mathbf{X}}$ is the density of (Y, δ) given $\mathbf{X}$, $f_{Y,\delta,\mathbf{X}}$ denotes the joint density of the data, and

$$R(t, \boldsymbol{x}) = \mathbb{E}(\mathbb{1}\{Y \geq t\} \mid \mathbf{X} = \boldsymbol{x}) f_{\mathbf{X}}(\boldsymbol{x}). \tag{4}$$

Here, $f_{\mathbf{X}}(\boldsymbol{x})$ denotes density of $\mathbf{X}$ with respect to the product measure between the Lebesgue measure and the counting measure on the support of the discrete covariates. See also (Efromovich, 2018, Ch.6, equation (6.2.5)) for an unconditional version of this representation of the hazard function. The function $R(t, \boldsymbol{x})$ is called the *natural nuisance function* in Efromovich (2024) and plays a key role in the minimax theory of nonparametric estimation of the conditional hazard function.

The representation (3) is a key relationship for building nonparametric estimator of the conditional hazard. Next, we show that similar representations can be derived in the presence of many other incompleteness mechanisms.

2.2 Beyond the right censoring case

In many applications a fraction of the right censored observed durations corresponds to subjects who will never experience the event of interest. In biostatistics such models are usually called *cure models*. Economists sometimes call such models *split population models*, while the reliability engineers refer to them as *limited-failure population* life models. The usual approach for such models is to suppose that the support of the duration T is $[0, \infty]$, while the right censoring C has no mass at infinity. See Patilea and Van Keilegom (2020) for a discussion. Let $\phi(\boldsymbol{x}) = \mathbb{P}(T < \infty \mid \boldsymbol{x}) < 1$ be the conditional probability of being susceptible to the event of interest given the covariates, and $1 - \phi(\boldsymbol{x})$ be the conditional probability of being cured.

Let $\lambda_{0,T|\mathbf{X}}(t \mid \boldsymbol{x})$ be the conditional hazard function of the susceptibles (often called the latency of the model), that are the individuals with a finite lifetime. By (Patilea and Van Keilegom, 2020, Equation (2.10)), in the case of the so-called mixture cure model, we get the following expression for the conditional hazard function of the susceptibles :

$$\begin{aligned}\lambda_{0,T|\mathbf{X}}(t \mid \boldsymbol{x}) &= \frac{f_{Y,\delta,\mathbf{X}}(t, 1, \boldsymbol{x})}{\{1 - H(t- \mid \boldsymbol{x})\} f_{\mathbf{X}}(\boldsymbol{x}) - \{1 - \phi(\boldsymbol{x})\} S_{C|\mathbf{X}}(t- \mid \boldsymbol{x}) f_{\mathbf{X}}(\boldsymbol{x})} \\ &= \frac{f_{Y,\delta,\mathbf{X}}(t, 1, \boldsymbol{x})}{R(t, \boldsymbol{x}) - \{1 - \phi(\boldsymbol{x})\} S_{C|\mathbf{X}}(t- \mid \boldsymbol{x}) f_{\mathbf{X}}(\boldsymbol{x})}, \qquad t \in [0, \tau(\boldsymbol{x})], \end{aligned} \tag{5}$$

with $R(t, \boldsymbol{x})$ defined in (4). It is worth noticing that $\{1 - \phi(\boldsymbol{x})\}$, $S_{C|\mathbf{X}}(t- \mid \boldsymbol{x})$ and $f_{\mathbf{X}}(\boldsymbol{x})$ can also be estimated by standard approaches. For example, by switching the roles of the variables T and C and redefining the censoring indicator as $1 - \delta$, C is right-censored by T with no cured subjects. Then, assuming C admits a conditional hazard, we can apply the approach in (3) to estimates its conditional hazard from which the estimator of $S_{C|\mathbf{X}}(t \mid \boldsymbol{x})$ can be derived.

Another possible extension of our approach is to consider *left-censoring* instead of right-censoring. See, for example, Gómez et al. (1992). In the left-censoring context, instead of duration of interest T, the observed variables are the duration $Y = T \vee C$ together with the indicator $\delta = \mathbb{1}\{T \geq C\}$. In this case, C is a left-censoring variable. By Assumption 1, we deduce

$$r_{T|\mathbf{X}}(t \mid \boldsymbol{x}) := \frac{f_{T|\mathbf{X}}(t \mid \boldsymbol{x})}{F_{T|\mathbf{X}}(t \mid \boldsymbol{x})} = \frac{f_{Y,\delta,\mathbf{X}}(t, 1, \boldsymbol{x})}{H(t \mid \boldsymbol{x}) f_{\mathbf{X}}(\boldsymbol{x})}. \tag{6}$$

Here, the function $r_{T|\mathbf{X}}$ is a conditional reverse time hazard function. Like the standard hazard function, it characterizes the conditional distribution of T given $\mathbf{X} = \boldsymbol{x}$ because the relationship $F_{T|\mathbf{X}}(t \mid \boldsymbol{x}) = \exp(-\int_t^\infty r_{T|\mathbf{X}}(s \mid \boldsymbol{x}) ds)$ holds true.

In addition to random right censoring, we can also consider a *left truncation* mechanism. To introduce the LTRC mechanism, that is *left-truncation in the presence of right-censoring*, we follow the lines in (Efromovich, 2018, Section 6.4). There is a hidden (latent) vector $(T^*, C^*, \mathbf{X}^*, L^*)$, where L^* is the truncation variable, T^* is the variable of interest, C^* is the right censoring variable, and $\mathbf{X}^*$ are the predictors. The independent latent sample is $(T_l^*, C_l^*, \mathbf{X}_l^*, L_l^*)$, $l \geq 1$, of $(T^*, C^*, \mathbf{X}^*, L^*)$. If $L_l^* > Y_l^* := T_l^* \wedge C_l^*$, the $l-$th realization in the latent sample is not observed, while if $L_l^* \leq T_l^* \wedge C_l^*$ the realization $(Y_l^*, \delta_l^*, \mathbf{X}_l^*, L_l^*)$, with $\delta_l^* = \mathbb{1}\{T_l^* \leq C_l^*\}$, becomes the $i-$th observation $(Y_i, \delta_i, \mathbf{X}_i, L_i)$ of the observed variables $(Y, \delta, \mathbf{X}, L)$. Assuming the mutual conditional independence $T^* \perp L^* \perp C^* \mid \mathbf{X}^*$, and that $\mathbb{P}(L^* \leq T^* \wedge C^* \mid \mathbf{X} = \boldsymbol{x}) \geq c > 0$, for some constant c, it can be shown that

$$\lambda_{T^*|\mathbf{X}^*}(t \mid \boldsymbol{x}) = \frac{f_{T^*|\mathbf{X}^*}(t \mid \boldsymbol{x})}{S_{T^*|\mathbf{X}^*}(t \mid \boldsymbol{x})} = \frac{f_{Y,\delta,\mathbf{X}}(t, 1, \boldsymbol{x})}{\mathbb{P}(L \leq t \leq Y \mid \mathbf{X} = \boldsymbol{x}) f_{\mathbf{X}}(\boldsymbol{x})}, \qquad \forall t \geq 0,\ \boldsymbol{x} \in \mathcal{X}. \tag{7}$$

See Efromovich (2018). Here, $\lambda_{T^*|\mathbf{X}^*}$, $S_{T^*|\mathbf{X}^*}$ and $f_{T^*|\mathbf{X}^*}$ are the conditional hazard, the conditional survival function and the conditional density of T^* given $\mathbf{X}^*$, respectively.

It is worth noting that both the truncation mechanism and the right censoring *with* a cure rate can be present at the same time. A conditional hazard function representation can be derived by combining (5) and (7). The details are given in Aurouet and Patilea (2025).

As a last example where our approach applies straightforwardly, let us consider the case studied by Patilea and Rolin (2006), where instead of the lifetime of interest T, one observes random copies of a duration $Y \geq 0$ and of a discrete variable $A \in \{0, 1, 2\}$ such that

$$Y < T \ \text{ if } \ A = 0, \quad Y = T \ \text{ if } \ A = 1 \quad \text{ and } \quad Y \geq T \ \text{ if } \ A = 2. \tag{8}$$

A classic example of this type of censoring, called *modified current status*, comes from the Stanford-Palo Alto Peer Counseling Program; see Hamburg et al. (1975). In this study, 193 high schools boys were asked 'When did you first use marijuana?'. The answers were the

exact ages ($A = 1$), 'I have used it but cannot recall just when the first time was' ($A = 2$) and 'I never used it' ($A = 0$). The variable of interest T is the age at first use of marijuana. It is worth noting that in (8) the limit case where the event $\{A = 2\}$ (resp. $\{A = 0\}$) has zero probability corresponds to the usual random right-censoring (resp. left-censoring) setup, while the case where the probability of the event $\{A = 1\}$ is null corresponds to the current status framework. In addition to Assumption 1, assume that $\Delta \perp (T, C) \mid \mathbf{X}$. Aurouet and Patilea (2025) show that, for any $\boldsymbol{x}$ in the support of $\mathbf{X}$, the following representation for the conditional hazard function of the duration of interest T holds true :

$$\lambda_{T|\mathbf{X}}(t \mid \boldsymbol{x}) = \frac{f_{Y,A,\mathbf{X}}(t, 1, \boldsymbol{x})}{[1 - H_1(t- \mid \boldsymbol{x}) + p(\boldsymbol{x})\{1 - H_0(t- \mid \boldsymbol{x})\}] f_{\mathbf{X}}(\boldsymbol{x})}, \tag{9}$$

where $f_{Y,A,\mathbf{X}}$ is the joint density of observed variables.

The list of possible models where the conditional hazard function has an explicit representation based on characteristics of the observed variables is even longer. See Aurouet and Patilea (2025). The approach can also be extended to bivariate durations under right censoring of each component. See Dabrowska (1988) and Efromovich (2022). However, there are other situations where the (conditional) hazard function does not seem to allow similar explicit representations as in the case of interval censoring, see Groeneboom (1996), or the left and right censoring model introduced by Turnbull (1974).

3 Concluding remarks

In this paper, we propose a new, general perspective on nonparametric regression for time-to-event data subject to censoring and/or truncation, allowing for cured individuals, applicable to univariate or bivariate lifetime variables. The approach is based on a representation of the conditional hazard function given the predictors as a ratio between a joint density and a conditional expectation. The joint density refers only to individuals for whom the time-to-event is observed. The representation can be established with continuous and discrete predictors and requires only mild, general identification conditions. For example, in the context of left or right censoring, the time-to-event needs to be conditionally independent of the censoring given the predictors.

Given the conditional hazard representation, we propose to estimate the numerator and the denominators by recursive kernel smoothing, which is well suited for online estimation. Uniform convergence in probability for a general class of such estimators is derived, and it is shown that they achieve optimal convergence rates. We show that the optimal bandwidth for a twice continuously differentiable joint density is proportional to the standard optimal bandwidth for the batch kernel density estimator. The expression of the proportionality constant is given, leading to a simple bandwidth rule for our conditional hazard estimator. The details are available in Aurouet and Patilea (2025). Simulation experiments will be presented, showing the good finite sample performance of our recursive estimator with right censoring. The method is applied to a real dataset of primary breast cancer.

References

Amico, M. and Van Keilegom, I. (2018). Cure models in survival analysis. *Annu. Rev. Stat. Appl.*, 5:311–345.

Aurouet, D. and Patilea, V. (2025). Continuously updated approach to estimating conditional lifetime distributions. *arXiv:2503.08356*.

Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *Ann. Stat.*, 16(4):1475–1489.

Efromovich, S. (2018). *Missing and modified data in nonparametric estimation*, volume 156 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL. With R examples.

Efromovich, S. (2022). Nonparametric bivariate density estimation for censored lifetimes. *Ann. Stat.*, 50(5):2767–2792.

Efromovich, S. (2024). Conditional hazard rate estimation for right censored data. *Bernoulli*, 30(3):2423 – 2449.

Gómez, G., Julià, O., and Utzet, F. (1992). Survival analysis for left censored data. In *Survival analysis: state of the art (Columbus, OH, 1991)*, volume 211 of *NATO Adv. Sci. Inst. Ser. E: Appl. Sci.*, pages 269–288. Kluwer Acad. Publ., Dordrecht.

Groeneboom, P. (1996). Lectures on inverse problems. In *Lectures on probability theory and statistics (Saint-Flour, 1994)*, volume 1648 of *Lecture Notes in Math.*, pages 67–164. Springer, Berlin.

Hamburg, B. A., Kraemer, H. C., and Jahnke, W. (1975). A hierarchy of drug use in adolescene: behavioral and attitudinal correlates of substantial drug use. *Amer. J. Psychiatry*, 132(11):1155–1163.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data (2nd edition)*. Springer-Verlag, New York.

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H., editors (2014). *Handbook of survival analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL.

Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis*. Statistics for Biology and Health. Springer, New York, third edition. A self-learning text.

Patilea, V. and Rolin, J.-M. (2006). Product-limit estimators of the survival function for two modified forms of current-status data. *Bernoulli*, 12(5):801–819.

Patilea, V. and Van Keilegom, I. (2020). A general approach for cure models in survival analysis. *Ann. Statist.*, 48(4):2323–2346.

Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.*, 69:169–173.

Sélection de variables pour données de survie censurées par intervalles

Ariane Bercu[1,a] & Agathe Guilloux[2,b] & Cécile Proust-Lima[3*,a] & Hélène Jacqmin-Gadda[4*,a]

[a] *Univ. Bordeaux, INSERM, BPH, U1219, F-33000 Bordeaux, France*
[b] *Inria Paris F-75015, INSERM, CRC, Univ. Sorbonne Paris, Sorbonne Université, F-75006 Paris, France*
[1]*ariane.bercu@u-bordeaux.fr* [2]*agathe.guilloux@inria.fr*
[3]*cecile.proust-lima@u-bordeaux.fr* [4]*helene.jacqmin-gadda@u-bordeaux.fr*
∗*co-derniers auteurs*

Résumé. Le modèle illness-death pour données censurées par intervalles permet d'étudier les facteurs de risque d'une maladie en présence d'un risque semi-compétitif de décès lorsque le diagnostic est réalisé uniquement à des visites discrètes, comme c'est le cas pour la survenue d'un syndrome de démence chez les personnes âgées à partir de données de cohortes. Dans ce travail, nous proposons une méthodologie de régularisation du modèle illness-death permettant la sélection de variables en grande dimension tout en traitant la censure par intervalles. Nous introduisons un critère d'estimation défini comme la vraisemblance du modèle prenant en compte la censure par intervalles et incluant un terme de pénalisation elastic-net pour chacune des trois transitions. Les paramètres de pénalisation des 3 transitions sont estimés par maximisation du BIC. Pour chaque jeu de paramètres de pénalisation, la vraisemblance pénalisée est maximisée par une procédure hybride de descente de gradient alternant une mise à jour des paramètres de régression et l'estimation des paramètres des fonctions de risque de base par l'algorithme de Marquardt. Les performances de cet algorithme ont été évaluées par simulations et comparées à l'approche par risques compétitifs plus communément utilisée qui néglige la censure par intervalles. Le modèle illness-death pour données censurées par intervalles a permis de sélectionner les variables pour chaque transition. A contrario, lorsque la censure par intervalles est négligée, la procédure sélectionne à tort des variables associées au décès pour la transition vers la démence. Les capacités prédictives du modèle illness-death sont supérieures à l'approche par risques compétitifs avec une plus faible erreur quadratique moyenne de prédiction de la démence. Ainsi, négliger la censure par intervalles de la démence peut amener à une sélection erronée de facteurs de risque de la démence et détériorer la prédiction des futurs cas de démence. La méthodologie est implémentée dans le packages R *HIDeM* et appliquée aux données de la cohorte en population des Trois Cités.

Mots-clés. Censure, multi-état, pénalisation, lasso, elastic-net, illness-death, démence.

Abstract. The illness-death model for interval-censored data accounts for the uncertainty on the time of dementia and the probability of having dementia until the last visit or death. In this work, we proposed a new regularised estimation procedure for a high dimensional illness-death model with interval censored data, that performs variable selection. Variable selection will allow us to increase predictive abilities and possibly identify new risk factors of dementia. We considered an hybrid algorithm maximising the regularised likelihood with

elastic-net penalty on the transitions to dementia and those to death with and without dementia. Our algorithm simultaneously estimates all three transitions' regression parameters while having different penalty parameters on each transition. The optimal penalty parameter is chosen via the bayesian information criteria. The performances of our algorithm were evaluated in simulations and compared to standard competing risk approach that neglects interval censoring. The illness-death model accounting for interval-censored data provided great predictions for the transition to dementia and good performances for variable selection. In comparison, the standard competing risk model showed worse predictive abilities and tended to select irrelevant variables for transition to dementia when associated with death. Neglecting interval censoring of dementia can deteriorate prediction of future cases of dementia and lead to a misleading selection of risk factors. Variable selection in illness-death model handling interval censoring data allows to accurately predict future cases of dementia while identifying risk factors of dementia among a large number of predictors. The regularized illness-death model for interval censored data is implemented in an R package *HIDeM*.

Keywords. Censoring, multi-state, regularisation, lasso, elastic-net, illness-death, dementia.

1 Introduction

Les modèles multi-états permettent d'analyser le processus d'évolution d'un sujet à travers différents états successifs d'une maladie [1]. Parmi eux, le modèle illness-death est un modèle multi-état où les sujets entrent sains, peuvent développer une maladie et décéder avec ou sans la maladie. Le changement d'un état à l'autre est défini par les intensités de transitions *sain-malade*, *malade-décédé* et *sain-décédé*. Dans les études de cohorte en population, le diagnostic de la démence est effectué lors des visites de suivi par un clinicien, ce qui entraîne une censure par intervalles du temps de survenue de la démence entre la visite de diagnostic et la visite précédente. De plus, lorsque le décès survient entre l'apparition de la démence et la visite suivante planifiée, la démence ne peut pas être diagnostiquée. Lorsque la méthode d'estimation ne tient pas compte de la censure par intervalles, le risque peut donc être sous-estimé.

Ainsi, la majorité des études sur la démence utilise un modèle pour risques compétitifs cause-specifique qui néglige la censure par intervalles [2, 3]. Le modèle cause-specifique modélise le temps de survenue de la démence en considérant le décès comme une censure. En particulier, les sujets non-diagnostiqués déments sont considérés comme non-déments jusqu'à leurs décès. Le temps de démence pour les sujets diagnostiqués est imputé au milieu de l'intervalle entre la visite de diagnostic et la visite précédente. Ces hypothèses peuvent amener à une sous-estimation de l'incidence de la démence et biaiser les estimation des associations des variables explicatives avec le risque de démence [4]. Joly et al.[5] ont proposé un algorithme d'estimation du modèle illness-death pour données censurées par intervalles. Pour les sujets ayant une démence, la contribution à la vraisemblance est la probabilité de développer une démence entre la visite de diagnostic et la visite précédente. Pour les sujets décédés sans diagnostic de démence, la contribution à la vraisemblance est la somme de la probabilité des deux trajectoires passant ou non par la démence. Alors que des approches de

sélection de variables ont été développées pour le modèle cause-spécifique à risques compétitifs standard [6, 7], aucune méthode de selection de variables n'est disponible pour le modèle illness-death avec censure par intervalles.

Il est établi que la survenue de la démence est associée à de multiples marqueurs socio-démographiques, de santé globale ainsi qu'aux capacités cognitives et aux volumes cérébraux [8, 9]. De ce fait, une méthodologie permettant de modéliser la survenue de la démence en prenant en compte la censure par intervalles et un grand nombre de prédicteurs semble être un outil statistique prometteur.

L'objectif de ce travail est d'étendre le modèle illness-death à la grande dimension des variables explicatives. Nous proposons une méthodologie de régularisation élastic-net du modèle illness-death pour faire de la sélection de variables en grande dimension tout en traitant la censure par intervalles. La capacité prédicitive et la pertinence de la sélection sont comparées à celles du modèle à risques compétitifs négligeant la censure par intervalles dans une étude de simulations. Finalement, le modèle illness-death en grande dimension est appliqué aux données de la cohorte populationelle des 3-Cités (3C) afin de sélectionner les principaux prédicteurs du risque de démence [10].

2 Le modèle

Le modèle illness-death est défini par un processus d'évolution entre trois états : sain (état 0), malade (état 1) et décès (état 2). Seuls les sujets sains (état 0) au début de l'étude sont inclus. Au cours du suivi, certains peuvent développer une maladie, 0→1, puis décéder, 1→2 ou directement décéder sans maladie 0→2. L'évolution d'un état à l'autre est déterminée par des intensités de transitions, pour $hl \in 01{,}02{,}12$,

$$\alpha_{hl}(t) = \lim_{\delta t \to 0} \frac{p_{hl}(t, t+\delta t)}{\delta t}$$

avec $p_{hl}(t, t+\delta t)$ la probabilité d'être dans l'état l au temps $t+\delta t$ sachant que le sujet se trouvait dans l'état h au temps t. Joly et al. ont proposé une intensité de transition proportionnelle de l'état h à l permettant d'inclure des facteurs de risques propre à chaque transition en supposant une proportionnalité des effets [11], comme dans un modèle de Cox classique. Ainsi, pour $hl \in 01{,}02{,}12$,

$$\alpha_{hl}(t|\theta_{hl}, \beta_{hl}, Z_{hl}) = \alpha_{0,hl}(\theta_{hl}, t) e^{\beta_{hl}^\top Z_{hl}}$$

où Z_{hl} est le vecteur des variables explicatives de la transition $h \to l$, β_{hl} les coefficients de régression et $\alpha_{0,hl}(\theta_{hl}, t)$ l'intensité de transition de base de la transition $h \to l$ au temps t avec θ_{hl} les coefficients associés. Dans un contexte de grande dimension, le nombre de coefficients de régression pour chaque transition, notés β_{hl}, augmente drastiquement, et une régularisation peut être appliquée. La log-vraisemblance régularisée du modèle illness-death est définie comme

$$l_{pen}(\theta, \beta) = l(\theta, \beta) - pen_{01}(\beta_{01}) - pen_{02}(\beta_{02}) - pen_{12}(\beta_{12})$$

où $\theta = (\theta_{01}, \theta_{02}, \theta_{12})$ est l'ensemble des paramètres des fonctions d'intensité de base, $\beta = (\beta_{01}, \beta_{02}, \beta_{12})$ est l'ensemble des paramètres associés aux variables explicatives, $l(\theta, \beta)$ est la

log-vraisemblance en (θ, β) et $pen_{hl}(\beta_{hl})$ correspond à la pénalisation de la transition $h \to l$. En présence de censure par intervalles, la vraisemblance fait intervenir tous les chemins de progression possibles pour chaque individu. Nous nous référons à Joly et al. [5] pour son écriture. Pour $hl \in$ 01,02,12, nous utilisons une combinaison convexe de pénalisation avec,

$$pen_{hl}(\beta_{hl}) = \lambda_{hl} a \|\beta_{hl}\|_1 + \lambda_{hl}(1-a)\|\beta_{hl}\|_2^2$$

communément appelée Elastic-Net, avec $\lambda_{hl} > 0$ le paramètre de pénalisation de la transition $h \to l$ et $0 \leq a \leq 1$. L'écriture de cette vraisemblance pénalisée permet d'estimer simultanément les effets des variables sur les trois transitions tout en appliquant une pénalisation propre à chaque transition. Cette spécificité est nécessaire pour permettre une sélection potentiellement différente des variables explicatives entre chaque transition.

3 Méthode d'estimation

Pour maximiser la log-vraisemblance pénalisée $l_{pen}(\theta, \beta)$ à paramètres de pénalisation fixés, nous avons utilisé un algorithme de descente de gradient de par coordonnées [6, 7] qui alterne les mises à jour des paramètres d'association pénalisés et des paramètres des fonctions d'intensité de base.

A l'itération k, pour des paramètres de pénalisation fixes, nous effectuons les étapes suivantes :

---

1. **Pour chaque élément p du vecteur β, mettre à jour $\widehat{\beta}_p^{(k)}$ pour $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(\boldsymbol{k-1})}$ et $\boldsymbol{\beta_{-p}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{-p}}^{(\boldsymbol{k-1})}$ d'après $\widetilde{l}_{pen}(\beta_p|\widehat{\boldsymbol{\theta}}^{(\boldsymbol{k-1})}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{-p}}^{(\boldsymbol{k-1})})$ définie comme**

$$\widetilde{l}_{pen}(\beta_p|\widehat{\boldsymbol{\theta}}^{(\boldsymbol{k-1})}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{-p}}^{(\boldsymbol{k-1})}) \;=\; \widetilde{l}(\beta_p|\widehat{\boldsymbol{\theta}}^{(\boldsymbol{k-1})}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{-p}}^{(\boldsymbol{k-1})}) - \underset{p\in 01}{pen}(\beta_p) - \underset{p\in 02}{pen}(\beta_p) - \underset{p\in 12}{pen}(\beta_p)$$

avec $\widetilde{l}(\beta_p|\widehat{\boldsymbol{\theta}}^{(\boldsymbol{k-1})}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{-p}}^{(\boldsymbol{k-1})})$ le développement de Taylor à l'ordre 2 au voisinage de $\widehat{\beta}_p^{(k-1)}$.

2. **Estimer $\widehat{\boldsymbol{\theta}}^{(\boldsymbol{k})}$ pour $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(\boldsymbol{k})}$ en maximisant $l(\boldsymbol{\theta}|\widehat{\boldsymbol{\beta}}^{(\boldsymbol{k})})$ avec un algorithme de Marquardt-Levenberg débutant à $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(\boldsymbol{k-1})}$** [12].

3. **Evaluer la convergence** définie comme une stabilisation des estimations des paramètres $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})$ et de la log-vraisemblance pénalisée l_{pen}.

Si les conditions de convergence ne sont pas atteintes, continuez vers l'itération $k+1$

Cet algorithme en deux étapes permet d'estimer nos paramètres de manière efficace tout en conservant l'objectif de pénalisation sur les paramètres de régression. Les paramètres de pénalisation optimaux $(\lambda_{01}, \lambda_{02}, \lambda_{12}, a)$ sont sélectionnés par minimisation du critère d'information bayésien (BIC).

4 Simulations

Pour évaluer cette méthodologie, nous avons simulé un suivi de cohorte populationnelle. La cohorte comprend 2000 sujets sains suivis pendant 18 années. Parmi eux, 600 sont diagnostiqués déments au cours du suivi, 1200 décèdent sans démence et 400 décèdent avec une démence. Nous avons simulé deux taux de censure par intervalles, définis comme le pourcentage de démences non diagnostiquées du fait d'un décès : un taux faible de 15% lorsque les visites sont planifiées tous les 2.5 ans et élevé à 30% lorsque les visites sont espacées de 4.5 ans. Pour chacune des transitions, 50 variables explicatives sont incluses mais seulement 15 d'entre elles sont associées à au moins une des transitions.

Nous avons évalué les performances de l'approche en terme de capacités prédictives, de sélection et d'estimation des coefficients des variables explicatives. Les capacités prédictives sont évaluées par l'erreur quadratique moyenne de probabilité d'avoir une démence avant un temps t où t est le temps minimal entre la démence, le décès ou la censure (MSEP). Pour obtenir une distribution du MSEP, nous avons effectué une sélection de variables sur 500 jeux de données d'entraînement, puis, à partir de ces estimations, nous avons calculé le MSEP sur 500 jeux de données tests. Nous avons comparé notre approche à celle du modèle oracle, qui connaitrait les variables réellement associées aux évènements, et au modèle à risques compétitifs négligeant la censure par intervalles. Dans la figure ci-dessous, nous constatons que la distribution du MSEP pour le modèle illness-death, estimé par vraisemblance penalisée en prenant en compte la censure par intervalles, est très proche du modèle oracle (qui n'inclut que les variables pertinentes) estimé par le maximum de vraisemblance en traitant la censure par intervalles. La régression pénalisée avec le modèle à risques compétitifs négligeant la censure par intervalles donne des MSEP nettement plus élevés, d'autant plus lorsque le taux de censure par intervalles augmente.
Concernant la sélection de variables, le modèle illness-death prenant en compte la censure par intervalles a permis de sélectionner presque exclusivement les variables associées à chaque transition. A contrario, le modèle à risques compétitifs négligeant la censure par intervalles a sélectionné des variables associées au décès après démence comme prédicteurs de la transition à la démence.

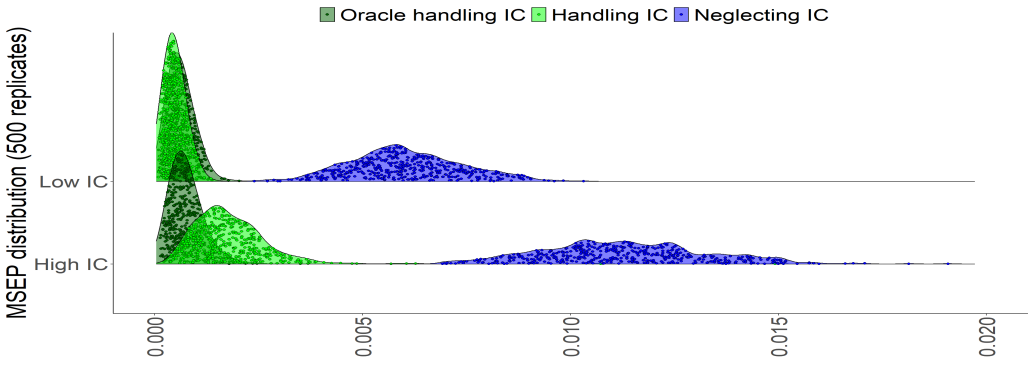


Figure 1: Distribution de l'erreur quadratique moyenne de prédiction de la démence en traitant la censure par intervalles (vert clair) et en la négligeant (bleu) (500 répliques).

5 Application

Le modèle a été appliquée aux données des 9031 participants de la cohorte des trois-Cités (3C) afin d'étudier de l'association entre de multiples volumes cérébraux et le risque de démence tout en prenant en compte les facteurs socio-démographiques et cognitifs.

References

1. Andersen PK, Abildstrom SZ, and Rosthøj S. Competing risks as a multi-state model. Statistical Methods in Medical Research 2002; 11:203–15
2. Andersen PK and Keiding N. Interpretability and importance of functionals in competing risks and multistate models. Statistics in Medicine 2012; 31:1074–88
3. Andersen PK, Geskus RB, De Witte T, and Putter H. Competing risks in epidemiology: possibilities and pitfalls. International Journal of Epidemiology 2012; 41:861–70
4. Leffondré K, Touraine C, Helmer C, and Joly P. Interval-censored time-to-event and competing risk with death: is the illness-death model more accurate than the Cox model? International Journal of Epidemiology 2013; 42:1177–86
5. Joly P, Commenges D, Helmer C, and Letenneur L. A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. Biostatistics 2002; 3:433–43
6. Simon N, Friedman J, Hastie T, and Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software 2011; 39
7. Friedman J, Hastie T, and Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 2010; 33
8. Stephan BCM, Tzourio C, Auriacombe S, Amieva H, Dufouil C, Alperovitch A, and Kurth T. Usefulness of data from magnetic resonance imaging to improve prediction of dementia: population based cohort study. BMJ 2015; 350(jun22 8):h2863–h2863
9. Hou XH, Feng L, Zhang C, Cao XP, Tan L, and Yu JT. Models for predicting risk of dementia: a systematic review. Journal of Neurology, Neurosurgery & Psychiatry 2019; 90:373–9
10. Group TCS. Vascular Factors and Risk of Dementia: Design of the Three-City Study and Baseline Characteristics of the Study Population. Neuroepidemiology 2003; 22:316–25
11. Touraine C, Helmer C, and Joly P. Predictions in an illness-death model. Statistical Methods in Medical Research 2016; 25:1452–70
12. Philipps V, Hejblum P. B, Prague M, Commenges D, and Proust-Lima C. Robust and Efficient Optimization Using a Marquardt-Levenberg Algorithm with R Package marqLevAlg. The R Journal 2021; 13:273

Apprentissage prédictif en analyse de survie par maximisation empirique de l'indice de concordance de Harrell

Stephan Clémençon [1], Florian Lamalle [2] & Anne Sabourin [3]

[1] *LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France et* stephan.clemencon@telecom-paris.fr
[2] *Groupe Renault et Université Paris Cité, CNRS, MAP5, F-75006 Paris, France et* florian.lamalle@renault.com
[3] *Université Paris Cité, CNRS, MAP5, F-75006 Paris, France et* anne.sabourin@math.cnrs.fr

Résumé. L'analyse de survie traite des durées avant la survenue d'un événement d'intérêt, souvent soumises à des données censurées. Ce travail se concentre sur l'élaboration d'une fonction de classement $f : \mathbb{R}^d \to \mathbb{R}_+$ destinée à prioriser des profils X selon la durée associée T, non totalement observée car censurée par une variable aléatoire C. Cette fonction doit maximiser l'indice de concordance de Harrell (C-index), définit comme $f \mapsto \mathbb{P}(f(X) < f(X') \mid T < T')$, qui mesure la capacité de f à ordonner correctement les durées.

Dans un cadre pratique, les données disponibles incluent des durées censurées, où seules des valeurs tronquées $\tilde{T} = \min(T, C)$ et un indicateur binaire $\delta = \mathbb{I}\{T \leq C\}$ sont observés. Pour traiter ce défi, nous proposons une version empirique corrigée du C-index adaptée à ces données. Nous démontrons également des bornes de vitesse d'apprentissage pour les maximisateurs de cet estimateur empirique, garantissant une performance plus robuste que les approches existantes.

Sur le plan théorique, notre approche repose sur la théorie des U-statistiques pour gérer les déviations dues au caractère par paires du C-index et aux biais causés par la censure. Nos résultats sont validés par des expérimentations numériques, confirmant leur pertinence dans des applications variées, notamment la maintenance prédictive et l'épidémiologie.

Mots-clés. Analyse de survie, censure à droite, indice de concordance, U-statistiques, maintenance prédictive.

Abstract. Survival analysis focuses on the durations before the occurrence of events of interest, often subject to censoring. This work focuses on designing a ranking function $f : \mathbb{R}^d \to \mathbb{R}_+$ to prioritize profiles X based on the associated duration T, which is not fully observed due to censoring by a random variable C. This function aims to maximize Harrell's concordance index (C-index), defined as $f \mapsto \mathbb{P}(f(X) < f(X') \mid T < T')$, which evaluates the ability of f to rank durations accurately.

In practical settings, the available data include censored durations, where only truncated values $\tilde{T} = \min(T, C)$ and a binary indicator $\delta = \mathbb{I}\{T \leq C\}$ are observed. To address this challenge, we propose a corrected empirical version of the C-index tailored to censored data. We also demonstrate learning rate bounds for the maximizers of this empirical estimator, ensuring more robust performance compared to existing approaches.

From a theoretical perspective, our approach builds on U-statistics theory to manage deviations caused by the pairwise nature of the C-index and the biases induced by censoring. Our results are validated through numerical experiments, confirming their relevance in various applications, including predictive maintenance and epidemiology.

Keywords. Survival analysis, right-censoring, concordance index, U-statistics, predictive maintenance.

1 Résumé long

L'analyse de survie est une branche cruciale des statistiques qui s'intéresse aux variables temporelles jusqu'à la survenue d'un événement d'intérêt, comme la défaillance d'un système mécanique ou le décès d'un patient, cf (Jones, 2011), (Meeker and Escobar, 2014), (Crowder et al., 1994), (Mann et al., 1974) ou (Kleinbaum and Klein, 2005) par exemple. Contrairement aux données classiques, les données de survie sont souvent censurées : pour certaines observations, l'événement n'est pas observé durant la période d'étude, ce qui complexifie leur traitement. Cette censure, généralement à droite (on observe uniquement le minimum entre le temps de l'événement T et un temps de censure C), nécessite des méthodologies spécifiques pour éviter des biais d'estimation (Nelson (2009), Kalbfleisch and Prentice (2011)). Les applications de l'analyse de survie s'étendent à des domaines variés : fiabilité industrielle (maintenance prédictive), finance (prédiction de défauts de crédit), médecine (pronostic de survie), ou gestion de la relation client (prédiction de désabonnements).

1.1 Défis et objectifs

L'objectif principal est souvent de prédire la probabilité de survie conditionnelle $S_T(t \mid X) = \mathbb{P}(T > t \mid X)$, où $X \in \mathbb{R}^d$ avec $d \geq 1$, représente des covariables explicatives (âge, caractéristiques techniques, etc.). Cependant, dans de nombreuses applications, il est plus utile de "seulement" classer les individus par ordre de risque (décès, défaillances par exemple) que de chercher à prédire exactement quand l'événement se produira. Par exemple, pour la maintenance prédictive, il s'agit de prioriser les interventions sur les équipements les plus susceptibles de tomber en panne. Ce classement repose sur l'index de concordance de Harrell (C-index), une mesure évaluant la capacité d'un modèle à ordonner correctement les paires d'individus selon leur temps jusqu'à l'événement. Le C-index, introduit par Harrell et al. (1982), est défini comme la probabilité conditionnelle

$$\mathcal{C}(f) = \mathbb{P}(f(X) < f(X') \mid T < T'), \tag{1}$$

où $f : \mathbb{R}^d \mapsto \mathbb{R}_+$ est une fonction de classement, et (X', T') est une copie indépendante de (X, T). Un C-index proche de 1 indique un classement presque parfait, tandis qu'une valeur de 0.5 correspond à un modèle aléatoire. L'estimation de ce critère à partir de données censurées pose deux défis majeurs :

- Le conditionnement de (1) : Le conditionnement par $T < T'$ n'est observable qu'indirectement à travers les réalisations du triplet $(X, \tilde{T}, \delta)$, et non via celles du couple (X, T).
- La nature combinatoire du critère : Le C-index implique des comparaisons par paires, conduisant à des estimateurs de type U-statistiques (de la Peña and Giné (1999)), complexes à analyser théoriquement.

1.2 Travaux Antérieurs et Limites

Plusieurs approches ont été proposées pour adapter le C-index aux données censurées : des méthodes naïves comme $f \mapsto \mathbb{P}(f(X) < f(X') \mid \tilde{T} < \tilde{T}', \delta = 1)$ utilisant des temps observés $\tilde{T}$ et T' afin d'éviter d'avoir à réecrire le critère (1) en fonction du triplet $(X, \tilde{T}, \delta)$ simplifiant l'estimation au prix d'un changement de cible, ce qui biaise l'interprétation du C-index (Schmid et al. (2015), Steck et al. (2007) et Pölsterl (2020)). Des méthodes de pondération par des inverses de probabilité de censure (IPCW) corrige le biais mais souffre d'instabilités numériques dues à la division par des estimateurs de survie $\widehat{S}_C$ de S_C, potentiellement proches de zéro (Uno et al. (2011)). Cette méthode se base sur la réecriture de (1)

$$\mathcal{C}(f) = 2\mathbb{E}\left[\frac{\delta\delta' \mathbb{I}\left\{f(X) < f(X')\right\} \mathbb{I}\left\{\tilde{T} < \tilde{T}'\right\}}{S_C(\tilde{T} \mid X) S_C(\tilde{T}' \mid X')}\right].$$

1.3 Contributions principales

Nous proposons une nouvelle approche pour l'estimation robuste du C-index et l'apprentissage de fonctions de classement f, avec des garanties théoriques et une mise en œuvre efficace. Pour cela, nous reformulons notre critère en exploitant les propriétés des fonctions de survie conditionnelles $S_T(t \mid X)$ et montrons que

$$\mathcal{C}(f) = -2\mathbb{E}\big[\mathbb{I}\{f(X) < f(X')\} \int S_T(t|X') S_T(dt|X)\big],$$

Cette expression évite les divisions instables et permet d'utiliser des estimateurs standards de $S_T(\cdot \mid X)$ (e.g., Kaplan-Meier ou Beran, Beran (1981)). À partir de cette réecrite du proposons donc l'estimateur suivant basé sur un échantillon d'entraînement $\{(X_i, \tilde{T}_i, \delta_i) : i = 1, \ldots, n\}$ composé de copies indépendantes de $(X, \tilde{T}, \delta)$,

$$\widehat{\mathcal{C}}(f) = -\frac{2}{n(n-1)} \sum_{i \neq j} \mathbb{I}\left\{f(X_i) < f(X_j)\right\} \int \widehat{S}_T(t \mid X_j) \widehat{S}_T(dt \mid X_i). \tag{2}$$

Nous établissons des bornes non-asymptotiques sur les déviations maximales $\sup_{f \in \mathcal{F}} \mathcal{C}(f) - \mathcal{C}(\hat{f})$ où $\hat{f} \in \operatorname{argmax}_{f \in \mathcal{F}} \hat{\mathcal{C}}(f)$ valables pour des classes de fonctions $\mathcal{F}$ de complexité VC contrôlées à l'aide de techniques de linéarisation et de techniques de décomposition orthogonales relatives à la théorie des U-processus (de la Peña and Giné (1999), Clémençon et al. (2006)). Nous établissons que la vitesse de convergence de l'indice de concordance (C-index)

empirique est de l'ordre de $\mathcal{O}_{\mathbb{P}}(\log(n)^{1/20}/n^{1/[5(d+4)]})$. Pour maximiser (2), nous remplaçons la fonctions indicatrice non-différentiable par des substituts concaves (e.g., logistique ou hinge loss) différentiables, permettant une optimisation par descente de gradient stochastique. Des expériences numériques sur des données simulées et réelles illustrent la robustesse de l'estimateur proposé face à la censure, contrairement aux méthodes IPCW.

1.4 Conclusion

Ces travaux combine rigueur théorique et applicabilité pratique, offrant un cadre robuste pour le classement en analyse de survie.

Bibliographie

References

Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical Report*.

Clémençon, S., Lugosi, G., and Vayatis, N. (2006). Ranking and empirical minimization of u-statistics.

Crowder, M., Kimber, A., Sweeting, T., and Smith, R. (1994). *Statistical Analysis of Reliability Data*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

de la Peña, V. and Giné, E. (1999). *Decoupling: From Dependence to Independence*. Probability and Its Applications. Springer New York.

Harrell, F., Califf, R., Pryor, D., Lee, K., and Rosati, R. (1982). Evaluating the yield of medical tests. *JAMA*, 247(18):2543—2546.

Jones, B. (2011). A class of three-level designs for definitive screening in the presence of second-order effects. *Journal of Quality Technology*, 43:1.

Kalbfleisch, J. and Prentice, R. (2011). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley.

Kleinbaum, D. and Klein, M. (2005). *Survival Analysis: A Self-Learning Text*. Statistics for biology and health. Springer.

Mann, N., Schafer, R., and Singpurwalla, N. (1974). *Methods for Statistical Analysis of Reliability and Life Data*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section. Wiley.

Meeker, W. and Escobar, L. (2014). *Statistical Methods for Reliability Data*. Wiley Series in Probability and Statistics. Wiley.

Nelson, W. B. (2009). *Accelerated testing: statistical models, test plans, and data analysis*. John Wiley & Sons.

Pölsterl, S. (2020). scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6.

Schmid, M., Wright, M., and Ziegler, A. (2015). On the use of harrell's c for clinical risk prediction via random survival forests.

Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P., and Raykar, V. C. (2007). On ranking in survival analysis: Bounds on the concordance index. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Uno, H., Cai, T., Pencina, M., D'Agostino, R., and Wei, L. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30:1105–17.

Données fonctionnelles

Optimal inference for the mean of random functions

Omar Kassi [1] & Valentin Patilea [2]

[1] *Ensai, CREST - UMR CNRS 9194, France; omar.kassi@ensai.fr*
[2] *Ensai, CREST - UMR CNRS 9194, France; valentin.patilea@ensai.fr*

Résumé. Les fonctions aléatoires constituent un outil de modélisation naturellement adapté à de nombreuses applications. Nous étudions l'inférence de la moyenne de fonctions aléatoires à valeurs réelles définies sur un hypercube. Les fonctions aléatoires indépendantes sont observées sur un sous-ensemble discret et aléatoire de points de design, éventuellement avec un bruit hétéroscédastique. Nous proposons un nouvel estimateur basé sur des séries de Fourier et atteignant la vitesse minimax. Nous prouvons une borne non-asymptotique pour l'approximation gaussienne de notre estimateur. Des bandes de confiance pour la fonction moyenne sont ensuite construites.

Mots-clés. analyse des données fonctionnelles, bandes de confidence, base de Fourier multivariée, l'operateur de *de La Vallée Poussin*, méthode de Stein

Abstract. Random functions are a natural modeling tool in many applications. We study inference for the mean of real-valued random functions defined on a hypercube. The independent random functions are observed on a discrete, random subset of design points, possibly with a heteroscedastic noise. New optimal rate estimators based on Fourier series are proposed. A non-asymptotic bound for the Gaussian approximation of our estimator is proved and inference is constructed.

Keywords. confidence bands, *de La Vallée Poussin* operator, functional data analysis, multivariate Fourier basis, Stein's method

1 Introduction

The modeling of random functions is a burgeoning field that deals with modern, complex data, such as those collected by sensors. It has proved to be successful in a wide-ranging number of applications, such as medical, economics and sports data. See for instance, the textbooks Ramsay and Silverman (2005), Horváth and Kokoszka (2012), Kokoszka and Reimherr (2017), among other references. Constructing the inference for the mean function from discretely sampled noisy functions arise naturally as a fundamental challenge. We here propose a novel approach for the estimation of mean functions defined on multidimensional domains, for which we derive a non-asymptotic Gaussian representation. Optimal rate pointwise confidence intervals and uniform confidence bands are derived.

2 The setup

Let $\mathcal{T} = [0,1]^D$, for some $D \geq 1$, be the domain of the real-valued random function X. The random copies of X, denoted by X_i, $i = 1, 2, .., N$, are referred to as trajectories or sample paths, and are supposed to be continuous. It is also assumed that the mean and the covariance functions of X, *i.e.,*

$$\mu(\mathbf{t}) = \mathbb{E}[X(\mathbf{t})], \qquad \gamma(\mathbf{s},\mathbf{t}) = \mathbb{E}\left[\{X(\mathbf{s}) - \mu(\mathbf{s})\}\{X(\mathbf{t}) - \mu(\mathbf{t})\}\right], \quad \text{for any } \mathbf{s},\mathbf{t} \in \mathcal{T},$$

exist and are continuous. The data are generated as follows. For each $1 \leq i \leq N$, and given a positive integer M_i, let $\mathbf{T}_{i,m} \in \mathcal{T}$, $1 \leq m \leq M_i$, be the observation design points for X_i. The data associated with the trajectory X_i consist of the pairs $(Y_{i,m}, \mathbf{T}_{i,m}) \in \mathbb{R} \times \mathcal{T}$ where

$$Y_{i,m} = X_i(\mathbf{T}_{i,m}) + \varepsilon_{i,m}, \quad \text{with} \quad \varepsilon_{i,m} = \sigma(\mathbf{T}_{i,m})e_{i,m}, \quad 1 \leq i \leq N,\ 1 \leq m \leq M_i,$$

with $e_{i,m}$ independent copies of a centered variable e with unit variance, and $\sigma(\cdot)$ is some unknown bounded function which takes into account possible heteroscedastic measurement errors. The design points $\mathbf{T}_{i,m}$ are assumed to be random vectors, independent copies of $\mathbf{T} \in \mathcal{T}$ which admits a density $f_{\mathbf{T}}$ with respect to the Lebesgue measure. This is the so-called *independent design* framework. Finally, the number of all data points $(Y_{i,m}, \mathbf{T}_{i,m})$ used to estimate the mean $\mu(\cdot)$ is $\overline{M} = M_1 + \cdots + M_N$. For simplicity, $M_1, \ldots, M_N$ are non-random.

Next, we impose the following regularity condition for the mean function.

Definition 1. *Let α and C be two positive numbers. The Hölder class $\Sigma(\alpha, C; D)$ is defined as the linear space of functions $g : \mathcal{T} \to \mathbb{R}$ for which all order $\lfloor\alpha\rfloor$ partial derivatives exist and, for any $\boldsymbol{\alpha} \in \mathbb{N}^D$ with $|\boldsymbol{\alpha}|_1 := \alpha_1 + \alpha_2 + \cdots + \alpha_D = \lfloor\alpha\rfloor$, it holds that*

$$|\partial^{\boldsymbol{\alpha}} g(\mathbf{t}) - \partial^{\boldsymbol{\alpha}} g(\mathbf{s})| \leq C\|\mathbf{t} - \mathbf{s}\|^{\alpha - \lfloor\alpha\rfloor}, \qquad \forall \mathbf{t}, \mathbf{s} \in \mathcal{T},$$

where $\partial^{\boldsymbol{\alpha}} = \partial_1^{\alpha_1}\partial_2^{\alpha_2}\ldots\partial_D^{\alpha_D}$ is the multi-index partial derivatives operator.

2.1 Multivariate Fourier series and de La Vallée Poussin operator

Let us recall that the Fourier basis on $[0,1]$ is defined as follows :

$$\phi_0(t) = 1, \quad \phi_{2k-1}(t) = \sqrt{2}\sin(2\pi kt), \quad \phi_{2k}(t) = \sqrt{2}\cos(2\pi kt), \quad t \in [0,1], \quad k = 1, 2, \ldots.$$

Given $\mathbf{k} = (k_1, k_2, \ldots, k_D) \in \mathbb{N}^D$, define $\phi_{\mathbf{k}}(\mathbf{t}) = \prod_{j=1}^D \phi_{k_j}(t_j)$, with $\mathbf{t} = (t_1, \ldots, t_D) \in \mathcal{T}$. The L-th partial sum of the Fourier series for the function μ is defined for any $\mathbf{t} \in \mathcal{T}$ as:

$$S_L(\mu, \mathbf{t}) = \sum_{|\mathbf{k}|_1 \leq 2L} a_{\mathbf{k}}\phi_{\mathbf{k}}(\mathbf{t}), \qquad \text{with Fourier coefficients} \quad a_{\mathbf{k}} = \int_{\mathcal{T}} \mu(\mathbf{t})\phi_{\mathbf{k}}(\mathbf{t})\,\mathrm{d}\mathbf{t},$$

where $|\mathbf{k}|_1 = k_1 + \cdots + k_D$. $S_L(\mu, \cdot)$ is also called a *triangular* partial sum. Next, the *de La Vallée Poussin* sum, defined as :

$$V_L(\mu, \mathbf{t}) = \frac{1}{L}\left\{S_L(\mu, \mathbf{t}) + S_{L+1}(\mu, \mathbf{t}) + \cdots + S_{2L-1}(\mu, \mathbf{t})\right\}, \quad \mathbf{t} \in \mathcal{T}.$$

Proposition 1. *Assume $\mu \in \Sigma(\alpha_\mu, C_\mu; D)$ with α_μ and C_μ are two positive numbers. Then, a positive constant $\mathcal{C}_\mu$ exists, depending only on α_μ, such that $|V_L(\mathbf{t}) - \mu(\mathbf{t})| \leq \mathcal{C}_\mu L^{-\alpha_\mu}$, $\forall \mathbf{t} \in \mathcal{T}$.*

Proposition 1 can be derived from Németh (2014, Theorem 2.2). For the univariate case ($D = 1$), see DeVore and Lorentz (1993, Ch 7) and Efromovich (1999, Proposition 2.4.2).

3 Fourier coefficients inference

In Kassi and Patilea (2025), the following assumptions are considered

Assumptions.

(H1) The $\mathbf{T}_{i,m}$'s are random copies of $\mathbf{T} \in \mathcal{T}$, and $\overline{M} := \sum_{i=1}^{N} M_i \geq 4$.

(H2) The vector $\mathbf{T} \in \mathbb{R}^D$ admits a density $f_{\mathbf{T}}$; constants C_0, C_1 exists such that $0 < C_0 \leq f_{\mathbf{T}} \leq C_1$.

(H3) The mean function $\mu(\cdot)$ belongs to the Hölder class $\Sigma(\alpha_\mu, C_\mu; D)$, for some $\alpha_\mu, C_\mu > 0$.

(H4) The curves X_i are independent sample paths of a second order process X with mean function μ and covariance function γ. Moreover, $\int_{\mathcal{T}} \gamma(\mathbf{t}, \mathbf{t}) \mathrm{d}\mathbf{t} < \infty$.

(H5) The heteroscedastic errors are $\varepsilon_{i,m} = \sigma(\mathbf{T}_{i,m}) e_{i,m}$ with $e_{i,m}$ independent copies of a centered variable e with unit variance.

(H6) The sample paths of X, the realizations of $\mathbf{T}$ and e are mutually independent.

(H7) The density $f_{\mathbf{T}}(\cdot)$ and the conditional variance function $\sigma^2(\cdot)$ of the errors are α_f-Hölder and $\alpha_\sigma-$Hölder continuous functions, respectively, for some $\alpha_f, \alpha_\sigma > 0$.

(H8) There exist $H_\gamma \in (0, 1]$ and $L_\gamma > 0$ such that, for any $\mathbf{s} \in \mathcal{T}$, the function $\mathbf{t} \mapsto \gamma(\mathbf{s}, \mathbf{t})$ is a $H_\gamma-$Hölder continuous function with Hölder constant L_γ.

In Kassi and Patilea (2025) an estimator $\widehat{\mu}(\cdot; L)$ of the mean function $\mu(\cdot)$ is constructed and it is shown that with the following optimal choice of the integer L

$$L^* = \left\lfloor C^* \overline{M}^{\,-1/(2\alpha_\mu + D)} \right\rfloor \qquad \text{with} \qquad C^* = \left\{ \frac{(2\alpha_\mu \mathcal{C}_\mu)(D+1)!}{D K_1 \varrho(D) (2^{2D+1} - 2^D)} \right\}^{1/(2\alpha_\mu + D)}, \tag{1}$$

a non-asymptotic risk bound is obtained under the previous assumptions:

$$\mathbb{E} \|\widehat{\mu}(\cdot, L^*) - \mu(\cdot)\|_2^2 \leq \left\{ \mathfrak{C}_\mu \overline{M}^{\,-\frac{2\alpha_\mu}{2\alpha_\mu + D}} + K_2 \sum_{i=1}^{N} \frac{M_i(M_i - 1)}{\overline{M}(\overline{M} - 1)} \right\} \times \{1 + o(1)\},$$

with $K_2 = \int_{\mathcal{T}} \gamma(\mathbf{t}, \mathbf{t}) \mathrm{d}\mathbf{t}$ and $\mathfrak{C}_\mu$ is some explicit constant depending on D, α_μ, $\mathcal{C}_\mu$ from Proposition 1, and $K_1 = \int_{\mathcal{T}} \{\gamma(\mathbf{t}, \mathbf{t}) + \sigma^2(\mathbf{t})\} / f_{\mathbf{t}}(\mathbf{t}) \mathrm{d}\mathbf{t}$.

The following notation is used below : for any $J \in \mathbb{N}$, let

$$\overline{J} := \text{card}\{\mathbf{k} \in \mathbb{N}^D : |\mathbf{k}|_1 \leq J\} \lesssim \max\{1, J^D\}. \tag{2}$$

The constant in the last inequality depends only on D. Moreover, $\forall J \in \mathbb{N}$, $(v_{\mathbf{k}})^{\top}_{|\mathbf{k}|_1 \leq J} \in \mathbb{R}^{\overline{J}}$ denotes the column matrix filled from top to bottom with the values $v_{\mathbf{k}}$, $\mathbf{k} = (k_1, k_2, \ldots, k_D)$, considered in lexicographic order.

We aim to construct inference for the pointwise error and the supremum error, defined as

$$\Delta(\mathbf{t}; L) = \widehat{\mu}(\mathbf{t}; L) - V_L(\mu; \mathbf{t}) \quad \text{and} \quad \Delta_\infty(L) = \sup_{\mathbf{t} \in \mathcal{T}} |\widehat{\mu}(\mathbf{t}; L) - V_L(\mu; \mathbf{t})|,$$

respectively, in particular with $L = L^*$ derived in (1). Both $\Delta(\cdot)$ and Δ_∞ depend on the random vector $(\widehat{a}_{\mathbf{k}} - a_{\mathbf{k}})^{\top}_{|\mathbf{k}|_1 \leq 4L^*-2}$, for which we will derive a non-asymptotic bound for the Gaussian approximation. More generally, the control of the Gaussian approximation error for vectors of Fourier coefficient estimators also facilitates the approximation of the distribution of other quantities of interest, such as $\|\widehat{\mu}^* - V_{L^*}(\mu, \cdot)\|_2$.

It is worth noting that

$$\mu^*(\mathbf{t}) - V_{L^*}(\mu; \mathbf{t}) = (\widehat{a}_{\mathbf{k}} - a_{\mathbf{k}})^{\top}_{|\mathbf{k}|_1 \leq 4L^*-2} \Phi_{L^*}(\mathbf{t}),$$

where, for $\mathbf{t} \in \mathcal{T}$ and $J \in \mathbb{N}^*$,

$$\Phi_J(\mathbf{t}) = \sum_{|\mathbf{k}|_1 \leq 2J} \phi_{\mathbf{k}}(\mathbf{t}) \mathbf{u}_{\mathbf{k}} + \sum_{\ell=1}^{J-1} \frac{J-\ell}{J} \Big\{ \sum_{|\mathbf{j}|_1 = 2J+2\ell-1} \phi_{\mathbf{j}}(\mathbf{t}) \mathbf{u}_{\mathbf{j}} + \sum_{|\mathbf{j}|_1 = 2J+2\ell} \phi_{\mathbf{j}}(\mathbf{t}) \mathbf{u}_{\mathbf{j}} \Big\},$$

and $(\mathbf{u}_{\mathbf{k}})_{0 \leq |\mathbf{k}|_1 \leq 4J-2}$ is the canonical basis of $\mathbb{R}^{\overline{4J-2}}$.

3.1 Berry-Esseen bounds by Stein's method

Let $d \in \mathbb{N}$ and consider the centered random vector $(\widehat{a}_{\mathbf{k}} - a_{\mathbf{k}})^{\top}_{|\mathbf{k}|_1 \leq d} \in \mathbb{R}^{\overline{d}}$. It can be shown that the leading term in the approximation of this vector is

$$W(d) = (V_{\mathbf{k}})^{\top}_{|\mathbf{k}|_1 \leq d} \in \mathbb{R}^{\overline{d}}, \qquad \text{with} \quad V_{\mathbf{k}} = \sum_{(i,m)} \omega_{i,m} \frac{\eta_{i,m}}{f_{\mathbf{T}}(\mathbf{T}_{i,m})} \phi_{\mathbf{k}}(\mathbf{T}_{i,m}),$$

$\eta_{i,m} = \{X_i - \mu\}(\mathbf{T}_{i,m}) + \sigma(\mathbf{T}_{i,m}) e_{i,m}$, and $\overline{d}$ defined according to (2). The quantity $V_{\mathbf{k}}$ is a variance term. Thanks to the control neighbors approach we propose, the bias term $(B_{\mathbf{k}} - a_{\mathbf{k}})$ is negligible and the distribution of $W(d)$ does not depend on the mean function which has to be estimated.

In order to derive the Gaussian approximation error bounds, we adopt Stein's method. More precisely, we follow the exchangeable pair approach in the multivariate setting, as developed by Reinert and Röllin (2009); see also Chatterjee and Meckes (2008).

Theorem 1. *Assume that (H1) to (H8) hold true. Moreover, for* $\mathbf{t}, \mathbf{s} \in \mathcal{T}$*, we assume that* $m_4(\mathbf{t}, \mathbf{s}) = \mathbb{E}\left[\{X-\mu\}^2(\mathbf{t})\{X-\mu\}^2(\mathbf{s})\right]$*, exists and the map* $(\mathbf{s}, \mathbf{t}) \mapsto m_4(\mathbf{t}, \mathbf{s})$ *is integrable. Let* $\Sigma = \mathbb{E}[W(d)W(d)^\top]$*. Then, if* Z *is a* $\overline{d}$*-dimensional standard Gaussian vector, we have for every three-times differentiable function* $h : \mathbb{R}^{\overline{d}} \to \mathbb{R}$*,*

$$\left|\mathbb{E}[h(W(d))] - \mathbb{E}[h(\Sigma^{1/2}Z)]\right| \leq \mathfrak{F}\overline{d}\left[\frac{|h|_3}{\overline{M}^{\,2}} + |h|_2\left\{\frac{1}{\overline{M}^{\,3}} + \frac{1}{\overline{M}^{\,4}}\sum_{i=1}^{N} M_i^2\right\}^{1/2}\right],$$

where $\mathfrak{F}$ *is a constant and*

$$|h|_2 := \sup_{p,q}\left\|\frac{\partial^2 h}{\partial x_p \partial x_q}\right\|_\infty, \qquad |h|_3 := \sup_{p,q,r}\left\|\frac{\partial^3 h}{\partial x_p \partial x_q \partial x_r}\right\|_\infty.$$

Theorem 1 allows to derive the normal approximation for several quantities that depend on Fourier coefficients, such as any linear combination. Let us define the following normalizing rates :

$$r_1(L) = \frac{\overline{M}}{L}, \qquad \text{and} \qquad r_2 = \frac{\overline{M}}{\sqrt{\sum_{1\leq i\leq N} M_i(M_i-1)}}.$$

Corollary 1. *The conditions of Theorem* (1) *are met, and* $\min\{r_1(L^*), r_2\} \to \infty$*. Then, for any vector* w *with bounded components of the same dimension as* $W(4L^*-2)$*, the variables*

$$\min\{r_1(L^*), r_2\}W(4L^*-2)^\top w \qquad \textit{and} \qquad \min\{r_1(L^*), r_2\}(\Sigma^{1/2}Z)^\top w,$$

have the same Gaussian asymptotic distribution, with L^* *as in* (1)*.*

Corollary 1 is obtained by applying Theorem 1 with the sub-class of three times differentiable functions h that can be written as $h(W(4L^*-2)) = g(W(4L^*-2)^\top w)$ with $g : \mathbb{R} \to \mathbb{R}$. Choosing $w = \Phi_{L^*}(\mathbf{t})$, for any $\mathbf{t} \in \mathcal{T}$, Corollary 1 implies that $\min\{r_1(L^*), r_2\}\Delta(\mathbf{t}; L^*)$ satisfies a Central Limit Theorem. For the supremum distance we have the following result.

Corollary 2. *Suppose the conditions of Corollary 1 hold true. Then, the random variables* $\min\{r_1(L^*), r_2\}\Delta_\infty(L^*)$ *and* $\min\{r_1(L^*), r_2\}\sup_{\mathbf{t}\in\mathcal{T}}\left|(\Sigma^{1/2}Z)^\top\Phi_{L^*}(\mathbf{t})\right|$ *have the same asymptotic distribution.*

For now, Corollaries 1 and 2 allow to construct confidence regions centered at *de La Vallée Poussin* sum $V_{L^*}(\mu, \mathbf{t})$. The following result proposes confidence regions centered at the mean function.

Corollary 3. *Suppose the conditions of Corollary 1 hold true. For any* $\vartheta(\overline{M}) \to \infty$ *such that, say,* $\log(\overline{M})/\vartheta(\overline{M}) \to \infty$ *such that*

$$L^\diamondsuit = \vartheta(\overline{M})\overline{M}^{\,\frac{1}{2\alpha_\mu+D}} \asymp \vartheta(\overline{M})L^*,$$

the quantities $\min\{r_1(L^\diamondsuit), r_2\}\left(\widehat{\mu}(\mathbf{t}, L^\diamondsuit) - \mu(\mathbf{t})\right)$ *and* $\min\{r_1(L^\diamondsuit), r_2\}(\Sigma^{1/2}Z)^\top\Phi_{L^\diamondsuit}(\mathbf{t})$*, have the same asymptotic distribution. Moreover,* $\min\{r_1(L^\diamondsuit), r_2\}\sup_{\mathbf{t}\in\mathcal{T}}|\widehat{\mu}(\mathbf{t}; L^\diamondsuit) - \mu(\mathbf{t})|$ *and* $\min\{r_1(L^\diamondsuit), r_2\}\sup_{\mathbf{t}\in\mathcal{T}}\left|(\Sigma^{1/2}Z)^\top\Phi_{L^\diamondsuit}(\mathbf{t})\right|$*, have the same asymptotic distribution.*

Remark 1. *It is worth noting that the Gaussian approximation given by the Corollary 3 does not depend on the mean function, which is the target of the inference. This pivotality property is due to our approach to the Fourier coefficients estimation using fast integration.*

Remark 2. *Corollary 3 holds for a sequence $\vartheta(\overline{M})$ which increases to infinity at an arbitrarily slow rate. The conclusion of Corollary 3 is also valid with $\vartheta(\overline{M}) \asymp 1$ as soon as $r_2/r_1(L^*) \to 0$, that is in the dense regime.*

References

Chatterjee, S. and Meckes, E. (2008). Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.*, 4:257–283.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation*, volume 303. Springer Science & Business Media.

Efromovich, S. (1999). *Nonparametric curve estimation*. Springer Series in Statistics. Springer-Verlag, New York. Methods, theory, and applications.

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*. Springer Series in Statistics. Springer, New York.

Kassi, O. and Patilea, V. (2025). Optimal inference for the mean of random functions. *Working Paper CREST Ensai*.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to functional data analysis*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL.

Németh, Z. (2014). On multivariate de La Vallée Poussin-type projection operators. *J. Approx. Theory*, 186:12–27.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition.

Reinert, G. and Röllin, A. (2009). Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Ann. Probab.*, 37(6):2150–2173.

Signal identification by local functional ANOVA

Rémi Mahmoud [1] & Ching-Fan Sheu [2] & David Causeur [3]

[1] *Institut Agro Rennes-Angers, Département Statistique et Informatique, UMR 6625 IRMAR CNRS remi.mahmoud@agrocampus-ouest.fr*
[2] *Institute of Education, National Cheng Kung University, csheu@ncku.edu.tw*
[3] *Institut Agro Rennes-Angers, Département Statistique et Informatique, UMR 6625 IRMAR CNRS david.causeur@agrocampus-ouest.fr*

Résumé. L'analyse des données fonctionnelles permet d'étudier des phénomènes à variations continues à partir de mesures discrètes. En particulier, l'analyse de la variance fonctionnelle peut être vue comme une transposition au cadre des données fonctionnelles des tests usuels de Fisher pour la comparaison de modèles linéaires. La problématique d'analyse de variance fonctionnelle la plus connue est probablement celle à un facteur du test de comparaison de courbes moyennes par groupes. Le test d'un effet sur l'ensemble de la plage d'observation des courbes est l'objet de nombreuses discussions, notamment sur la manière optimale d'agréger les statistiques de tests ponctuelles pour calculer une statistique de test global optimale et sur la gestion de la dépendance entre celles-ci. La détection par ce type de test d'un signal d'association entre des courbes moyennes et une ou plusieurs variables explicatives ouvre la voie à la question de l'identification de ce signal d'association, en d'autres termes à la recherche du support sur lequel le signal d'association est non-nul.

Dans de nombreux domaines d'application, cette question est abordée par les méthodes de tests multiples. Or, d'une part, ces méthodes ne prennent en général pas en compte la nature continue du signal d'association, ce qui peut conduire à identifier des temps de mesures isolés, dont l'interprétation est impossible. D'autre part, les méthodes de contrôle du nombre de faux positifs sont en général trop conservatrices, ce qui conduit souvent à n'identifier aucun intervalle support du signal d'association alors que le test global a conclu à l'existence de celui-ci à l'échelle de la plage d'observation. Ce problème se pose notamment en agronomie, où il est crucial d'identifier les périodes temporelles au cours desquelles des variations biologiques notables surviennent en fonction de covariables environnementales.

L'approche hybride que nous proposons pour l'identification des intervalles supports du signal d'association repose sur le F-test généralisé proposé par Causeur et al., 2020 pour l'analyse de la variance fonctionnelle de courbes d'électroencéphalogrammes dans des dispositifs expérimentaux impliquant plusieurs variables explicatives. À partir de l'application locale de ce test sur des intervalles de même amplitudes dont le centre parcourt la plage d'observation des courbes, la méthode vise à identifier la plus grande portion du signal dans laquelle le signal d'association n'est pas significatif. L'amplitude de l'intervalle glissant fait le compromis entre une procédure de tests multiples (intervalles d'un seul point) et un test global (intervalle couvrant toute la plage, fANOVA).

Après avoir détaillé l'algorithme et discuté des choix méthodologiques, nous illustrons son application sur des données simulées ainsi que sur des données réelles portant sur l'étude des

liens entre les émissions de mycotoxines par un champignon phytopathogène et l'excédent hydrique mesuré régulièrement au cours du temps pendant la croissance de la plante. L'application aux données réelles met en évidence la capacité de la méthode à identifier des périodes clés propices à l'émission de mycotoxines.

Mots-clés. ANOVA fonctionnelle, Test global, Identification d'un signal.

Abstract. Functional Data Analysis (FDA) focuses on studying time-continuous variations using discretized observations. Analogous to the standard F-test used to compare linear models when the response is scalar, Functional Analysis of Variance (fANOVA) provides a global test for assessing the effect of covariates on functional data. A well-known application of fANOVA is the comparison of mean curves in a one-way design. A central challenge in this context is the construction of a global test across the entire time domain, particularly regarding the aggregation of individual test statistics into an optimal global test statistic and the handling of temporal dependence across these statistics. Detecting a significant association between mean curves and one or more explanatory variables opens up the complementary objective of identifying time intervals where this association occurs.

In many cases, this identification problem is addressed using multiple testing procedures. However, these methods are not designed to account for the intrinsic functional nature of the association signal, leading to the detection of spurious, isolated significant time points. Additionally, the control of false positives tends to be overly conservative, frequently resulting in no significant intervals being identified, even when the global test indicates the presence of a non-zero association. This issue is particularly relevant in agricultural studies, where it is crucial to pinpoint time intervals during which biological variations are linked to environmental covariates.

To address this limitation, we propose a hybrid approach for identifying the temporal support of the association signal. Our method builds upon the generalized functional F-test introduced by Causeur et al., 2020 for fANOVA in electroencephalogram (EEG) studies involving multiple explanatory variables. The approach aims to identify the largest portion of the time domain where the association signal is not significant. This is achieved by applying the fANOVA test locally over intervals of fixed amplitude, with centers regularly distributed across the entire time frame. The amplitude of the sliding interval strikes a balance between a multiple testing procedure (intervals of a single point) and a global test (interval covering the entire range, fANOVA).

Following a detailed presentation of the algorithm, with a focus on key methodological choices, we assess its performance on both simulated and real data. Specifically, we analyse the emission of mycotoxins by a phytopathogenic fungus in relation to water excess, measured at regular intervals throughout plant growth. This case study demonstrates the method's ability to identify key time intervals that favour mycotoxin emissions, showcasing its potential for applications in agricultural and environmental research.

Keywords. Functional ANOVA, Global test, Signal identification

1 Functional Analysis of Variance

Functional data consist of observations of continuous curves measured on a discretized grid of points. The increasing availability of instruments and techniques — such as sensors, and spectroscopy — has led to a surge in high-throughput curve data, posing new challenges for statistical methodologies in Functional Data Analysis (e.g., Ramsay and Silverman, 2002, Ferraty, 2006). A notable challenge arises in neuroscience, where Event-Related Potentials (ERP) designs study the effects of experimental covariates on brain activity measured by Electroencephalography time-locked to an event of interest. This motivated Causeur et al., 2020 to propose a Functional Analysis of Variance (fANOVA) test, extending the standard F-test used for comparing linear models to accommodate functional responses.

Let $(Y(t))_{t\in\mathcal{T}}$ denote a functional response variable defined on a time domain $\mathcal{T}$. Let $x = (x_1, \ldots, x_p)'$ stand for a p-vector of time-independent explanatory variables. For all $t \in \mathcal{T}$, the following linear model is assumed:

$$Y(t) \quad = \quad \beta_0(t) + \beta_1(t)x_1 + \ldots + \beta_p(t)x_p + \varepsilon(t), \tag{1}$$

where $\beta_0(t)$ is the pointwise intercept parameter at time t and $\beta(t) = (\beta_1(t), \ldots, \beta_p(t))'$ is the p-vector of pointwise regression parameters at time t.

Analogously to the standard model comparison issue in linear model, the objective of fANOVA is to test general null hypotheses of the form H_0: $R\beta(t) = 0$, for all $t \in \mathcal{T}$, where R is a $k \times p$ matrix defining k linear contrasts and $\beta(t) = (\beta_1(t), \ldots, \beta_p(t))'$. For instance, in one-way designs where the only explanatory variable is a two-group covariate, the primary contrast is typically the mean difference curve. The fANOVA test proposed in Causeur et al., 2020 is designed to handle arbitrarily complex null hypotheses that may involve many explanatory variables. Rejecting H_0 implies that the association signal $t \mapsto \beta(t)$ is nonzero at some point, which is why this issue is often referred to as signal detection or global testing.

Model (1) is typically supplemented with specific assumptions regarding time-continuous variables. A common assumption is that the functions $t \mapsto \beta_j(t),\ j = 0, 1, \ldots, p$, exhibit a certain degree of regularity over time. This smoothness can be captured using linear decompositions on basis functions, such as B-splines. In this case, if $\phi(t) = (\phi_1(t), \ldots, \phi_q(t))$ represents the values of q basis functions $\phi_j(t)$, $j = 1, \ldots, q$ at time t, then $\beta(t) = \phi'(t)B$, where B is a $q \times p$ vector of basis functions coefficients. In the following, the signal identification method will be presented without this smoothing option. Finally, we assume $\varepsilon(t) \sim \mathcal{N}(0, \sigma^2(t))$ with the autocorrelation function $\rho(t, t') = \mathrm{cor}(\varepsilon(t), \varepsilon(t'))$. The functional responses $Y(t)$ are observed on a discrete time grid $t_1, \ldots, t_m$. For each $t \in \mathcal{T}$, model (1) corresponds to a standard linear model for a scalar response variable. Thus, pointwise F-test statistics $(F_{t_1}, \ldots, F_{t_m})$ can be used to test the null hypotheses H_{0,t_j}: $R\beta(t_j) = 0,\ j = 1, \ldots m$. Most fANOVA test statistics rely on simple aggregations – such as sums, weighted sums or maximum – of the pointwise F-statistics $F_{t_j},\ j = 1, \ldots, m$. Surprisingly, the time dependence between these pointwise F-statistics is often overlooked in the aggregation process (Górecki and Smaga, 2019). To ensure proper control of the type I error rate under arbitrary time dependence, the null distribution of the global test statistic is however estimated using random permutation techniques.

To account for the strong temporal dependence typically observed in functional data, the fANOVA test proposed by Causeur et al., 2020 addresses this dependence by constructing global test statistics as sums of fully or partially decorrelated pointwise test statistics. Partial decorrelation is achieved through linear operators derived from a k-factor approximation of the inverse correlation matrix of the pointwise test statistics. The number of factors, k, serves as a tuning parameter, ranging from 0 (no whitening) to the rank of the residual covariance matrix of model (1) (full whitening).

While Causeur et al., 2020 introduces a strategy for selecting the optimal number of factors, we propose an enhanced version of the test. In this new approach, a global omnibus test statistic is derived by aggregating the fANOVA test statistics computed for different choices of k, thereby incorporating information across multiple levels of decorrelation.

2 Signal identification by local fANOVA

When a global functional F-statistic detects a significant association signal over the entire time frame, it is natural to further investigate which specific segments contribute most to this finding. However, multiple testing procedures applied to the pointwise F-tests $F_{t_j},\ j = 1, \ldots, m$ often struggle to identify meaningful significant intervals. This limitation arises from two key factors: the strong time dependence between pointwise test statistics and the regularity of the association signal, which multiple testing fails to account for. As a result, these methods tend to either miss significant regions or detect spurious isolated time points.

Instead of attempting to identify significant time points, we propose a local functional ANOVA testing procedure that takes the opposite approach, *i.e* identifying the largest set of time intervals where no significant effect is detected using the functional ANOVA test introduced in Section 1. This procedure ensures consistency with the global fANOVA test. If the global test is not significant, then the largest set of time points where no significant effect is found spans the entire time frame, meaning no significant time points are identified. Conversely, if the global test is significant, then the largest non-significant set is necessarily smaller, ensuring coherence between local and global analyses.

To determine the maximal set of non-significant time points, we propose the following approach (Algorithm 1). First, we define a grid of regularly spaced time points across the entire time domain and test for the presence of an association signal within their neighbourhoods using the functional F-test introduced earlier. The neighborhood amplitude acts as a hyper-parameter, whose choice will be discussed during the presentation. If no significant intervals are found, the algorithm terminates, concluding no significant findings. Otherwise, a second step, backward elimination, sequentially removes significant intervals from the list until the fANOVA test outside these intervals becomes significant.

Algorithm 1 Signal identification

Input: l (window size), I (number of intervals), α (critical p-value threshold)
Compute interval centers $\mathcal{C} := (C_i)_{1 \leq i \leq I}$ with $C_i := 1 + (i-1) \cdot \frac{T-1}{I-1}$
for $i = 1$ to I **do**
    Define interval $J_i := \{t_j : |t_j - C_i| \text{ among the } l \text{ smallest}\}$
    Compute $p_i \leftarrow$ fANOVA p-value for J_i
end for
if $\min_i(p_i) \geq \alpha$ **then**
    return No interval identified
else
    Rank p_i values as $\tilde{p} := \{p_{(1)}, \ldots, p_{(S)}\}$ where $p_{(s)} < \alpha$ for $s \leq S$
    Initialize $\tilde{I}_1 \leftarrow I_1$, $s \leftarrow 1$, and $q_s \leftarrow 1$
    while $s \leq S \ \wedge \ q_s \geq \alpha$ **do**
        $s \leftarrow s + 1$
        Perform fANOVA on $T \setminus \tilde{I}_s$
        $q_s \leftarrow$ associated p-value
        if $q_s < \alpha$ **then**
            return $\bigcup_{s=1}^{S} I_s$ (significant intervals)
        else
            $\tilde{I}_s \leftarrow \tilde{I}_s \cup I_s$
        end if
    end while
end if

2.1 Application of the algorithm

2.1.1 Simulation

We test our method on simulated data, exploring parameters such as the number of individuals ($n \in \{100, 1000, 5000\}$), the temporal resolution ($T \in \{50, 500, 2000\}$), the presence of a temporal dependance (yes / no). For each of the combination we generate 100 datasets and compare the results to other methods ((Fused) LASSO, Tibshirani et al., 2005).

2.1.2 Application of the method on real-world data

Context: Fusarium Head Blight is a cereal disease that causes significant damage. The fungi responsible emit mycotoxins, such as Nivalenol (NIV), which pose risks to human and animal health. When contamination exceeds legal thresholds, wheat is devalued. Using a dataset from Arvalis (technical research institute) with data from 320 farms, we analyzed mycotoxin threshold exceedances alongside climatic time series (rainfall, evapotranspiration, global radiation, etc.). We applied our method to detect signals in water excess time series (P - ETP, mm) around flowering, linked to NIV threshold exceedances. The multiple testing procedure (pointwise F-tests) found no differences between curves ($p = 0.44$, Benjamini-Hochberg correction), while the fANOVA test introduced in Section 1 was significant ($p = 0$) but did not pinpoint specific segments responsible for this difference.

The two intervals identified (Fig. 1, $[-675°Cd, -350°Cd]$ et $[175°Cd, 525°Cd]$) correspond to a pre-flowering period where the moisture conditions are favorable for the growth of the

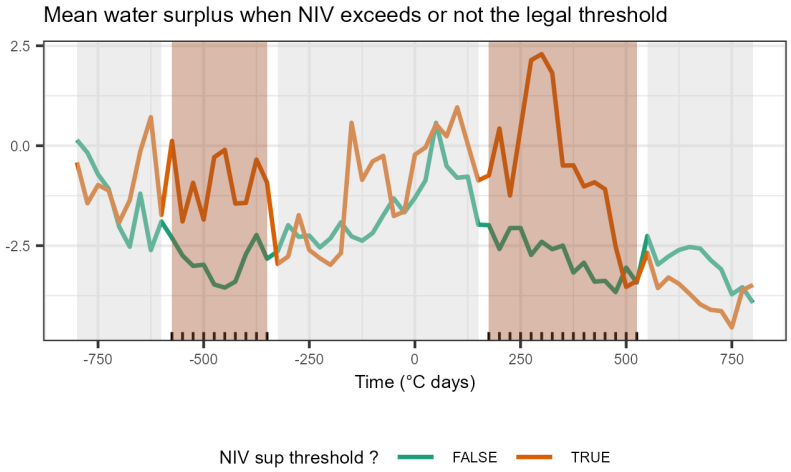


Figure 1: Application of the method to real-world data: Mean water surplus curves over time for cases exceeding (orange) or not (blue) the NIV legal threshold. The grey zone indicates time intervals with no significant effect (functional ANOVA), while the orange zone represents the complementary intervals likely causing the difference.

fungus before infection, and a post-flowering period where moisture around the spikes could promote the development of Fusarium as well as mycotoxin emission (Cowger et al., 2009).

3 Discussion and perspectives

The algorithm presented in this communication will be made available as an R package. A suitable criterion for selecting hyperparameters (l, I) still needs to be developed.

Références

Causeur, D., Sheu, C.-F., Perthame, E., & Rufini, F. (2020). A functional generalized f-test for signal detection with applications to event-related potentials significance analysis. *Biometrics*, *76*(1), 246–256. https://doi.org/10.1111/biom.13118

Cowger, C., Patton-Özkurt, J., Brown-Guedira, G., & Perugini, L. (2009). Post-anthesis moisture increased fusarium head blight and deoxynivalenol levels in north carolina winter wheat [PMID: 19271972]. *Phytopathology®*, *99*(4), 320–327. https://doi.org/10.1094/PHYTO-99-4-0320

Ferraty, F. (2006). *Nonparametric functional data analysis*. Springer. https://doi.org/https://doi.org/10.1007/0-387-36620-2

Górecki, T., & Smaga, Ł. (2019). Fdanova: An r software package for analysis of variance for univariate and multivariate functional data. *Computational Statistics*, *34*(2), 571–597. https://doi.org/10.1007/s00180-018-0842-7

Ramsay, J. O., & Silverman, B. W. (2002). *Applied functional data analysis: Methods and case studies*. Springer. https://doi.org/https://doi.org/10.1007/b98888

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(1), 91–108.

Prédiction du futur de données fonctionnelles à partir de covariables

Selman Sezgin [1], Julien Jacques [2], Kahina Mokrani [3] & Sylvain Allio [4]

[1] *Orange Innovation / Université de Lyon 2, Université Claude Bernard Lyon 1, ERIC, France, selman.sezgin@orange.com*
[2] *Université Lumière Lyon 2, Université Claude Bernard Lyon 1, ERIC, France, Julien.Jacques@univ-lyon2.fr*
[3] *Orange Innovation, France, kahina.mokrani@orange.com*
[4] *Orange Innovation, France, sylvain.allio@orange.com*

Résumé. Le présent travail aborde un problème de prédiction en analyse de données fonctionnelles. On considère la prédiction du futur d'une fonction Y à partir de l'observation de son passé ainsi que de celui de covariables fonctionnelles. Plus précisément, à partir d'une collection i.i.d. d'observations de Y et des covariables, nous développons un modèle qui, étant donnés le début de Y et le début des covariables, prédit le futur de Y. Notre méthode repose sur la décomposition du futur de Y dans une base de fonctions, et sur la prédiction des coefficients d'expansion par une approche de régression scalaire-sur-fonction. Les performances de notre modèle sont évaluées sur des données fonctionnelles synthétiques. Les résultats expérimentaux montrent que notre modèle permet de construire des prédictions plus lisses et plus précises que les méthodes traditionnelles existants.

Mots-clés. Analyse de données fonctionnelles, prédiction de données longitudinales, décomposition en fonctions de base.

Abstract. Functional data analysis is a branch of statistics where data are treated as smooth functions rather than finite-dimensional vectors. The functional approach allows for a more accurate representation of dynamic and continuous phenomena and has numerous real-world applications, such as in healthcare and finance. In this work, we address the problem of predicting a function Y based on functional covariates. More specifically, given an i.i.d. collection of observations of Y and the covariates, we develop a model that, given the beginning of Y and the beginning of the covariates, predicts the future of Y. Our method relies on decomposing the future of Y into a basis of functions and predicting the expansion coefficients using a scalar-on-function regression approach. The performance of our model is evaluated on both synthetic and real functional data. Experimental results show that our model provides smoother and more accurate predictions than traditional existing methods.

Keywords. Functional data analysis, longitudinal data prediction, functional basis expansion.

1 Introduction

L'analyse de données fonctionnelles est une branche active des statistiques qui traitent les données sous la forme de courbes lisses plutôt que de vecteurs de dimension finie. Le traitement des données par analyse fonctionnelle a de nombreux domaines d'application comme dans la santé [14], la finance [7] ou la biologie [5]. Dans ce cadre, contrairement à une approche d'apprentissage statistique traditionnelle, une donnée ne correspond pas à la réalisation d'un vecteur aléatoire de dimension finie, mais à une fonction lisse à domaine continue appartenant à un espace de dimension infinie, ce qui permet de refléter plus fidèlement des phénomènes dynamiques continus. Le livre [10] est une référence clé pour une introduction à l'analyse de données fonctionnelles. Les domaines de recherche sont variés et incluent la régression [13], le lissage [2], la classification [12] et le clustering [9]. Les données fonctionnelles sont souvent representées sous la forme de données dites *longitudinales*, c'est-à-dire où un (ou plusieurs) phénomène est observé sur un intervalle temporel pour plusieurs instances. Par exemple, une étude clinique pourrait suivre l'évolution dans le temps de paramètres physiologiques (poids, pression artérielle, température corporelle etc.) d'un ensemble d'individus. Dans le présent travail, nous étudions un problème de prédiction pour des données fonctionnelles longitudinales. Plus spécifiquement, nous considérons un ensemble de trajectoires d'une fonction d'intérêt, que l'on désignera par Y, et des trajectoires conjointes de covariables fonctionnelles. Notre but est de construire un modèle fonctionnel qui, à partir du début de la courbe de Y et des courbes des covariables, prédit la suite de la courbe de Y. Plus formellement, on considère une collection i.i.d. de données fonctionnelles

$$\Big\{Y_i(t), X_{1,i}(t), \ldots, X_{p,i}(t); t \in [0, T+H]\Big\}_{i=1}^{n} \tag{1}$$

où le $i-$ème échantillon contient un fonction $Y_i(t)$ et des covariables $X_{i,\ell}(t), \ell = 1, \ldots, p$, définies sur l'intervalle $[0, T+H]$, avec T et H strictement positifs. On considère ensuite un nouvel échantillon $Y_{n+1}(t), X_{n+1,\ell}(t), \ell = 1, \ldots, p$ où les fonctions ne sont observées que jusqu'au temps T et sont inconnues entre les temps T et $T+H$. Le but de ce travail est de faire une prédiction lisse de la suite de $Y_{n+1}(t)$ sur $[T, T+H]$ à partir de son observation passée sur $[0, T]$, de l'observation passée des covariables et du comportement général des fonctions sur les n échantillons complets initiaux (1). Le problème étudié est illustré dans la figure 1.

Nous proposons un modèle fonctionnel en trois étapes :

1. La fonction Y sur l'intervalle futur est décomposée dans une base de fonctions,
2. Un modèle fonctionnel paramétrique est optimisé pour estimer les coefficients d'expansion du futur de Y à partir du début de la courbe de Y et des covariables,
3. L'inférence pour la prédiction de Y sur le nouvel $(n+1)-$ème échantillon est faite à partir des paramètres optimisées et de l'observation passée des fonctions.

Les parties suivantes développent notre approche.

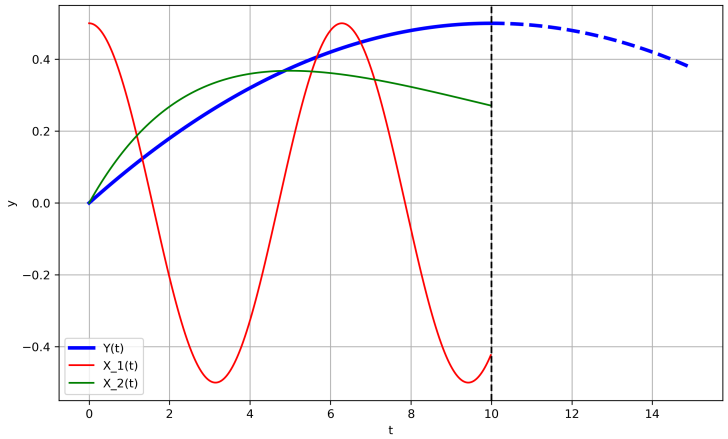


FIGURE 1 – Illustration du problème étudié. On cherche à prédire le futur de $Y(t)$ entre $T = 10$ et $T + H = 15$ à partir du passé de $Y(t)$ et du passé des covariables $X_1(t)$ et $X_2(t)$.

2 Modèle fonctionnel

Une hypothèse classique en analyse de données fonctionnelles est que les fonctions sous-jacentes aux observations appartiennent à l'espace de Hilbert des fonctions de carré intégrable, qui est un espace de dimension infinie. Pour réduire la dimensionalité du problème et faciliter le traitement statistique, il est courant de décomposer les données dans des bases de fonctions [10]. Des exemples classiques de bases sont les bases de Fourier pour les phénomènes périodiques [8], les ondelettes [3] et les splines qui sont largement utilisées en pratique pour représenter des fonctions non périodiques [1]. À travers la décomposition dans une base, l'information contenue dans une fonction à domaine continue est encodée par un nombre restreint de coefficients. Notre modèle repose sur l'idée clé que, étant donnée une base de fonctions, prédire le futur de Y revient à prédire ses coefficients dans cette base.

Notons $Y_i^{(1)}(t)$ et $Y_i^{(2)}(t)$ la restriction de $Y_i(t)$ sur l'intervalle $[0, T]$ et $[T, T + H]$ respectivement. De même, on introduit les notations $X_{i,\ell}^{(1)}(t)$ et $X_{i,\ell}^{(2)}(t)$ pour les covariables. Fixons une base de fonctions $\{\psi_k : k \in [\![1, K]\!]\}$ définies sur $[T, T + H]$. Alors la décomposition de $Y_i^{(2)}(t)$ dans cette base s'écrit

$$Y_i^{(2)}(t) = \sum_{k=1}^{K} b_{i,k}\psi_k(t) + \varepsilon_i(t), \tag{2}$$

où $\varepsilon_i(t)$ est une fonction résiduelle définie comme la composante de $Y_i^{(2)}(t)$ qui est linéairement indépendantes de la base $\{\psi_k\}$. En pratique, $\varepsilon_i(t)$ correspond au bruit des mesures ou a des comportements indésirables de la fonction. Par hypothèse du problème, les coefficients d'expansion dans la décomposition (2) sont calculables pour $i = 1, \ldots, n$ mais inconnus pour $i = n + 1$. Nous proposons alors de prédire les coefficients $\{\hat{b}_{n+1,k} : k \in [\![1, K]\!]\}$ puis de

reconstruire la suite de $Y_{n+1}^{(2)}(t)$ avec

$$\hat{Y}_{n+1}^{(2)}(t) = \sum_{k=1}^{K} \hat{b}_{n+1,k}\psi_k(t). \tag{3}$$

Nous traitons le calcul des $\{\hat{b}_{n+1,k}\}$ comme un ensemble de problèmes de régression scalaire-sur-fonction indépendants. Pour chaque $k = 1, \ldots, K$, les échantillons

$$\left\{(Y_i^{(1)}(t), X_{1,i}^{(1)}(t), \ldots, X_{p,i}^{(1)}(t)), b_{k,i}\right\}_{i=1}^{n} \tag{4}$$

induisent un problème de régression scalaire-sur-fonction où le coefficient $b_{k,i}$ est estimé à partir des prédicteurs fonctionnels $Y_i^{(1)}(t), X_{\ell,i}^{(1)}(t)$, $\ell = 1, \ldots, p$. Le modèle fonctionnel linéaire standard pour résoudre ce problème est le suivant [11] :

$$\hat{b}_{k,i} = \int_{u=0}^{T} Y_i^{(1)}(u)\beta_{k,0}(u)du + \sum_{\ell=1}^{p}\int_{u=0}^{T} X_{i,\ell}^{(1)}(u)\beta_{k,\ell}(u)du. \tag{5}$$

Les fonctions $\beta_{k,\ell}(t), \ell = 0, \ldots, p$ sont des paramètres définies sur $[T, T+H]$ et de carré intégrable. À l'aide des intégrales dans (5), ces paramètres permettent d'extraire l'information utile des fonctions en entrée pour la prédiction des coefficients d'expansion du futur de Y. En injectant (5) dans (2), notre modèle fonctionnel s'écrit :

$$\hat{Y}_i^{(2)}(t) = \sum_{k=1}^{K}\left(\int_{u=0}^{T} Y_i^{(1)}(u)\beta_{k,0}(u)du + \sum_{\ell=1}^{p}\int_{u=0}^{T} X_{i,\ell}^{(1)}(u)\beta_{k,\ell}(u)du\right)\psi_k(t). \tag{6}$$

En particulier, la courbe recherchée $\hat{Y}_{n+1}^{(2)}(t)$ peut être prédite par inférence à partir des paramètres $\{\beta_{k,\ell}\}$ estimés sur les n premiers échantillons et des courbes observées $Y_{n+1}^{(1)}(t)$ et $X_{\ell,n+1}^{(1)}(t)$, $\ell = 1, \ldots, p$.

3 Optimisation

En adoptant le principe classique de minimisation du risque empirique [4], l'optimisation du modèle (6) repose sur le calcul des paramètres fonctionnels $\{\beta_{k,\ell}\}$ de manière à minimiser l'erreur de prédiction sur une collection i.i.d. de données. Jusqu'à présent, nous avons considérés nos données fonctionnelles comme des courbes observées de manière continue dans le temps. Mais en pratique les fonctions sont observées sur un ensemble discret de points. Supposons que Y ainsi que les covariables sont mesurées aux temps $t_1 < \ldots < t_{m_1}$ sur l'intervalle $[0, T]$ et aux instants $t_{m_1+1} < \ldots < t_{m_1+m_2}$ sur $[T, T+H]$. L'optimisation des paramètres $\{\beta_{k,\ell}\}$ repose sur la minimisation de l'erreur quadratique suivante :

$$\min_{\{\beta_{k,\ell}\}} \quad \sum_{i=1}^{n}\sum_{j=m_1+1}^{m_1+m_2}\left(Y_i^{(2)}(t_j) - \hat{Y}_i^{(2)}(t_j)\right)^2. \tag{7}$$

Remarquons que dans le présent travail, la base de prédiction $\{\psi_k\}$ dans laquelle la suite de Y est estimée n'est pas optimisée. Pour une fonction Y générale, une base de B-splines est un choix convenable. La taille de la base est un hyperparamètre qui peut par exemple être choisi par validation croisée. Le problème (7) est difficile à résoudre directement : d'abord les paramètres recherchés $\{\beta_{k,\ell}\}$ appartiennent à un espace de dimension infinie, puis le modèle (6) ne reflète pas la nature discrète des données fonctionnelles en pratique. Pour surmonter ces difficultés, nous développons une formulation en dimension finie du problème (7). Pour ce faire, nous décomposons dans des bases fonctionnelles les fonctions qui interviennent dans le modèle (6), à savoir $Y_i^{(1)}(t)$, les covariables $X_{i,\ell}^{(1)}(t)$ ainsi que les paramètres $\{\beta_{k,\ell}\}$:

$$Y_i^{(1)}(t) = \sum_{j=1}^{h_0} c_{0,i,j}\theta_{0,j}(t); \quad X_{i,\ell}^{(1)}(t) = \sum_{j=1}^{h_\ell} c_{\ell,i,j}\theta_{\ell,j}(t); \quad \beta_{k,\ell}(t) = \sum_{j=1}^{q_\ell} b_{k,\ell,j}B_{\ell,j}(t) \tag{8}$$

Puis en injectant ces décompositions dans le modèle (6) et après quelques manipulations algébriques, on obtient une représentation en dimension finie de (6) qui s'écrit :

$$\hat{Y}_i^{(2)}(t) = \boldsymbol{\xi}_i(t)^\top \mathbf{b}. \tag{9}$$

avec $\mathbf{b}$ un vecteur qui encode l'ensemble des paramètres $\{\beta_{k,\ell}\}$ et un vecteur à dépendance temporelle $\boldsymbol{\xi}_i(t)$ qui encode les fonctions $Y_i^{(1)}(t)$ et $X_{i,\ell}^{(1)}(t), \ell = 1, \ldots, p$. Avec l'écriture (9), le problème (7) est réduit à un problème des moindres carré ordinaire :

$$\min_{\mathbf{b}} \quad \sum_{i=1}^{n} \sum_{j=m_1+1}^{m_1+m_2} \left(Y_i^{(2)}(t_j) - \boldsymbol{\xi}_i(t_j)^\top \mathbf{b} \right)^2. \tag{10}$$

En arrangeant les vecteurs ligne $\boldsymbol{\xi}_i(t_j)^\top$ dans une matrice $\boldsymbol{R}$ et les scalaires $Y_i^{(2)}(t_j)$ dans un vecteur $\boldsymbol{y}$, le problème (10) se réécrit de manière plus conventionnelles :

$$\min_{\mathbf{b}} \quad \|\boldsymbol{y} - \boldsymbol{R}\mathbf{b}\|^2. \tag{11}$$

Les coefficients optimaux $\mathbf{b}^*$ sont obtenus via l'inverse (ou la pseudo-inverse) de la matrice de Gram $\boldsymbol{R}^\top \boldsymbol{R}$.

4 Résultats expérimentaux

Notre modèle a été évaluée sur des données fonctionnelles synthétiques. Dans nos expériences, les fonctions sont mesurées toutes les secondes aux temps

$$t_1 = 1, , \ldots, t_{m_1} = T,\ t_{m_1+1} = T+1, \ldots, t_{m_1+m_2} = T+H \tag{12}$$

On suppose que $T+H = 300$ (i.e. 5 minutes) et que la fonction $Y(t)$ ne dépend que d'une seule covariable $X(t)$. Notons $G(t; \mu, \sigma) = e^{-(t-\mu)^2/(2\sigma)^2}$ la densité gaussienne (non normalisée). Les réalisations de $X(t)$ et $Y(t)$ sont générées à partir des équations suivantes :

$$X(t) = \sum_{u=1}^{10} c_u G(t; \mu_u, \sigma); \quad Y(t) = X(t) + A\sin(2\pi f t) \tag{13}$$

Modèle	$T = 240$, $H = 60$ ($\tau = 80\%$)	$T = 180$, $H = 120$ ($\tau = 60\%$)	$T = 120$, $H = 180$ ($\tau = 40\%$)
Lin	40.2 (59.5)	20.2 (19.1)	1.52 (0.08)
Ridge	0.403 (0.0083)	0.608 (0.00641)	0.69 (0.00715)
Lasso	0.402 (0.00683)	0.619 (0.00763)	0.702 (0.00892)
XGB	0.461 (0.0112)	0.658 (0.00691)	0.734 (0.0118)
Notre modèle	0.369 (0.00641)	0.602 (0.015)	0.691 (0.0105)

TABLE 1 – Performances de prédiction pour trois valeurs de τ. La moyenne et l'écart-type de la RMSE sont calculés sur les $n_{\text{simu}} = 30$ simulations.

où $c_u \sim \mathcal{U}[-1, 1]$, $\mu_1 < \ldots < \mu_{10}$ sont équidistants sur l'intervalle $[0, T + H]$, $\sigma \sim \mathcal{U}[20, 50]$, $A \sim \mathcal{U}[0.3, 0.7]$ et $f = 10^{-2}$. Pour chaque simulation, on génère $n_{\text{train}} = 300$ échantillons d'entraînement et $n_{\text{test}} = 2000$ échantillons de test. Un bruit gaussien de SNR = 15 dB est ajouté aux valeurs de $Y(t)$ et $X(t)$. Pour rappel, pour un SNR donné, le bruit gaussien est défini par $\varepsilon(t) \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ avec $\text{SNR} = 10 \log_{10}(P/\sigma_\varepsilon^2)$ où P est la puissance de la fonction, c'est-à-dire sa moyenne quadratique.

Le vecteur de paramètre $\mathbf{b}$ du modèle (9) est optimisé sur les échantillons d'entraînement puis l'inférence consiste à calculer les vecteurs $[\hat{Y}_i(t_{m_1+1}), \ldots, \hat{Y}_i(t_{m_1+m_2})]$ pour $i = 1, \ldots, n_{\text{test}}$. Des bases de B-spline cubiques ont été utilisées pour représenter les fonctions dans (8).

La métrique de performance du modèle est la RMSE (Root Mean Square Error) sur les échantillons de test :

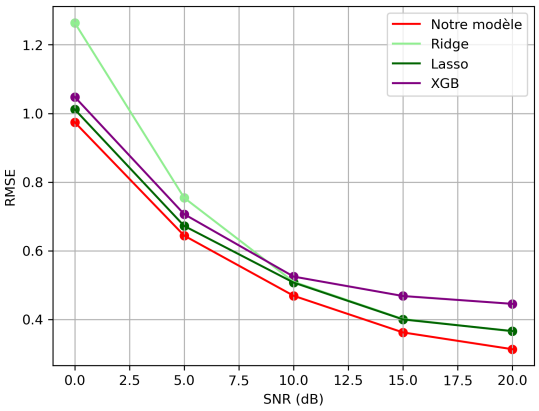
$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}} m_2} \sum_{i=1}^{n_{\text{test}}} \sum_{j=m_1+1}^{m_1+m_2} \left(Y_i(t_j) - \hat{Y}_i(t_j) \right)^2} \tag{14}$$

Le problème étudié peut être résolu par une approche discrète avec une régression de $\mathbb{R}^{2m_1}$ (passé de $Y(t)$ et $X(t)$) vers $\mathbb{R}^{m_2}$ (futur de $Y(t)$). La régression linéaire (Lin), la régression ridge (Ridge), la régression lasso (Lasso) et XGBoost [6] (XGB) ont été choisis comme modèles compétiteurs.

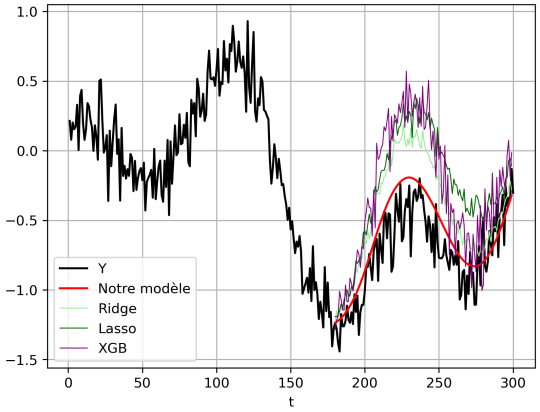
La première expérience consiste à évaluer l'impact des valeurs de T et H sur la précision des prédictions, sachant que la longueur totale $T + H = 300$ est fixée. Soit $\tau = \frac{T}{T+H} \times 100\%$ la portion de courbe observée. En général, la tâche de prédiction est d'autant plus difficile que τ diminue. Le tableau 1 synthétise les performances des modèles pour plusieurs valeurs de τ avec $n_{\text{simu}} = 30$ simulations.

La deuxième expérience consiste à fixer $T = 240$ et $H = 60$ et faire varier la force du bruit appliqué aux données. En général, la tâche de prédiction est d'autant plus difficile que le SNR est faible. Le graphique 2a compare les performances des modèles en fonction du niveau de bruit injecté, sur une seule simulation des données (des résultats moyens sur plusieurs simulation seront présentées lors de la conférence). Enfin, la figure 2b est un exemple d'inférence des modèles pour une simulation des données.

Les résultats expérimentaux dans 1 et 2a suggèrent que les prédictions de notre modèle sont plus précises que les modèles compétiteurs. Une meilleure robustesse vis-à-vis de la



(a) Performances de prédiction pour plusieurs niveaux de bruit.



(b) Exemple d'inférence sur un échantillon de test.

FIGURE 2 – Test de robustesse des modèles vis-à-vis du bruit et exemple d'inférence. La régression linéaire a été omise pour plus de visibilité car elle présente des erreurs de prédiction significativement supérieures.

fenêtre de prédiction et du bruit sont observées. De plus, le graphique 2b montre que notre modèle produit une prédiction lisse, ce qui a l'avantage d'une part de refléter la nature fonctionnelle des données, d'autre part d'avoir des prédictions plus simples à interpréter visuellement. Cela s'explique par le fait que la courbe prédite est par construction exprimée dans une base de fonctions (voir 3) tandis que les modèles compétiteurs sont discrets.

5 Conclusion

Nous avons développé un modèle fonctionnel pour prédire la suite d'une fonction à partir de son passé et du passé de covariables fonctionnelles. Notre approche repose sur la décomposition du futur de la fonction dans une base fonctionnelle appropriée, puis sur la prédiction des coefficients dans cette base. Des paramètres fonctionnels sont introduits et un problème des moindres carré ordinaire est formulé. Le modèle intègre de manière naturelle la fonction à prédire et les covariables. Les résultats expérimentaux montrent que l'approche fonctionnelle produit des prédictions lisses, plus précises et plus robustes que les modèles traditionnels existants. La suite de ce travail consistera à appliquer notre modèle dans un cas pratique avec des données réelles.

Références

[1] Ana M AGUILERA et MC AGUILERA-MORILLO. "Comparative study of different B-spline approaches for functional data". In : *Mathematical and Computer Modelling* 58.7-8 (2013), p. 1568-1579.

[2] Wan Anis Farhah Wan AMIR, Md Yushalify MISRO et Mohd Hafiz MOHD. "Flexible functional data smoothing and optimization using beta spline". In : *AIMS Mathematics* 9.9 (2024), p. 23158-23181.

[3] Anestis ANTONIADIS et al. "Clustering functional data using wavelets". In : *International Journal of Wavelets, Multiresolution and Information Processing* 11.01 (2013), p. 1350003.

[4] Francis BACH. *Learning theory from first principles*. MIT press, 2024.

[5] Veerabhadran BALADANDAYUTHAPANI et al. "Bayesian Hierarchical Spatially Correlated Functional Data Analysis with Application to Colon Carcinogenesis". In : *Biometrics* 64.1 (fév. 2008), p. 64-73. ISSN : 0006-341X. DOI : 10.1111/j.1541-0420.2007.00846.x.

[6] Tianqi CHEN et Carlos GUESTRIN. "XGBoost : A Scalable Tree Boosting System". In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA : Association for Computing Machinery, 2016, p. 785-794. ISBN : 9781450342322. DOI : 10.1145/2939672.2939785. URL : https://doi.org/10.1145/2939672.2939785.

[7] Sonali DAS et al. "The effect of global crises on stock market correlations : Evidence from scalar regressions via functional data analysis". In : *Structural Change and Economic Dynamics* 50 (2019), p. 132-147.

[8] Clara HAPP et Sonja GREVEN. "Multivariate functional principal component analysis for data observed on different (dimensional) domains". In : *Journal of the American Statistical Association* 113.522 (2018), p. 649-659.

[9] Julien JACQUES et Cristian PREDA. "Functional data clustering : a survey". In : *Advances in Data Analysis and Classification* 8 (2014), p. 231-255.

[10] James O. RAMSAY et Bernard Walter SILVERMAN. *Functional Data Analysis*. Springer, 2005. ISBN : 9780387400808. URL : http://www.worldcat.org/isbn/9780387400808.

[11] Philip T REISS et al. "Methods for scalar-on-function regression". In : *International Statistical Review* 85.2 (2017), p. 228-249.

[12] Fabrice ROSSI et Nathalie VILLA. "Support vector machine for functional data classification". In : *Neurocomputing* 69.7-9 (2006), p. 730-742.

[13] Jean Steve TAMO TCHOMGUI et al. "A mixture of experts regression model for functional response with functional covariates". In : *Statistics and Computing* 34.5 (juill. 2024). ISSN : 0960-3174. DOI : 10.1007/s11222-024-10455-z. URL : https://doi.org/10.1007/s11222-024-10455-z.

[14] Yuyan YI et al. "Classification of EEG signals : An interpretable approach using functional data analysis". In : *Journal of Neuroscience Methods* 376 (2022), p. 109609. ISSN : 0165-0270. DOI : https://doi.org/10.1016/j.jneumeth.2022.109609. URL : https://www.sciencedirect.com/science/article/pii/S0165027022001364.

Nouvelle approche de scan spatial pour données fonctionnelles : applications économiques et climatiques

Zaineb Smida,[1] Thibault Laurent,[2] & Lionel Cucala [2]

[1] *Univ Lyon, INSA Lyon, UJM, UCBL, ECL, CNRS UMR 5208, ICJ, F-69621, France, zaineb.smida@insa-lyon.fr*

[2] *Toulouse School of Economics, CNRS, France, Thibault.Laurent@tse-fr.eu* [3] *Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, France, lionel.cucala@umontpellier.fr*

Résumé. Une méthode de balayage pour des données fonctionnelles indexées dans l'espace a été développée. La statistique de balayage est dérivée de la statistique de test de Hotelling pour les données fonctionnelles, étendant ainsi les statistiques de balayage spatiales gaussiennes univariées et multivariées. Nos simulations montrent que cette nouvelle méthode surpasse systématiquement les techniques existantes pour détecter et localiser des clusters spatiaux. Elle a été appliquée à deux types de données réelles : des données économiques sur le taux de chômage afin d'identifier des clusters spatiaux sur les régions en Espagne, et des données climatiques afin de détecter des régions où le changement climatique opère différemment, en Grande-Bretagne.

Mots-clés. Détection de clusters, données fonctionnelles, test de Hotelling T^2, statistique de balayage spatiale.

Abstract. A scan method for functional data indexed in space has been developed. The scan statistic is derived from the Hotelling test statistic for functional data, extending the univariate and multivariate Gaussian spatial scan statistics. This method consistently outperforms existing techniques in detecting and locating spatial clusters, as demonstrated through simulations. It has been applied to two types of real data: economic data in order to identify spatial clusters of abnormal unemployment rates in Spain and climatic data in order to detect unusual climate change patterns in Great Britain.

Keywords. Cluster detection, Functional data, Hotelling T^2 test, Spatial Scan statistic.

1 Introduction

Les méthodes de détection d'agrégats (clusters) spatiaux ont fait l'objet de nombreuses études visant à développer des outils capables d'identifier des regroupements de sites présentant des comportements distincts. L'une des techniques les plus couramment utilisées pour détecter ces clusters est la méthode de balayage (scan) spatiale, introduite pour la première fois par Naus (1963).

Le domaine des méthodes de scan spatial a connu des avancées significatives grâce aux travaux de Kulldorff (1997), qui a introduit l'utilisation de fenêtres circulaires de tailles variables pour balayer la zone d'étude et identifier le cluster le plus probable à l'aide d'un test du rapport de vraisemblance. Il a utilisé soit le modèle de Bernoulli (Kulldorff & Nagarwalla 1995), soit le modèle de Poisson (Kulldorff 1997), et a évalué la significativité statistique des agrégats à l'aide d'une procédure de Monte Carlo. Ces avancées ont conduit à de nombreuses études (Huang et al. 2007, Kulldorff et al. 2009, Cucala 2014, 2016). Parfois, de nombreuses variables continues doivent être analysées simultanément. Plusieurs statistiques de scan multivariés en été introduites (Kulldorff et al. 2007, Cucala 2017, Cucala et al. 2019).

Les données étant de plus en plus mesurées en temps continu ou quasi-continu, des méthodes d'analyse des données fonctionnelles ont été développées (Ramsay & Silverman 2005). Dans le cadre des statistiques de scan spatial, Smida et al. (2022) et Frévent et al. (2021) ont proposé des statistiques adaptées à ce type de données. Ce travail introduit une nouvelle statistique de scan spatial pour données fonctionnelles, basée sur le test de Hotelling (Joseph et al. 2015, Horváth et al. 2013). Cette approche étend les travaux de Kulldorff et al. (2009), Cucala et al. (2017), offrant un outil robuste pour identifier les clusters.

2 Méthodologie

2.1 Introduction de la statistique de scan HFSS

On considère une variable aléatoire X dans l'espace fonctionnel $\chi = L^2(\mathcal{T}, \mathbb{R})$ et des observations $X_1, \ldots, X_n$ en n localisations distinctes $s_1, \ldots, s_n$ dans un domaine $D \subset \mathbb{R}^2$. L'objectif est de détecter un agrégat de marques inhabituelles, soit une zone $Z \subset D$ où les marques fonctionnelles diffèrent de celles observées ailleurs. Dans ce travail, nous considérons les clusters circulaires de Kulldorff (1997). L'ensemble des clusters potentiels $\mathcal{S}$ est défini par :

$$\mathcal{S} = \{D_{i,j}, 1 \leq i, j \leq n\},$$

où $D_{i,j}$ est le disque centré en s_i et passant par s_j, avec $s_j \in D_{i,j}$.

Par la suite, supposons que $X_1, \ldots, X_n$ soient des observations indépendantes de X. Soit $Z \in \mathcal{S}$ un agrégat potentiel de taille $n_Z = \sum_{i=1}^{n} \mathbb{1}(s_i \in Z)$ et Z^c son complément de taille $n_{Z^c} = n - n_Z$. Ainsi, dans le cadre de la détection de clusters, l'hypothèse nulle H_0 (absence de cluster) peut être formulée comme suit : $H_0 : \forall Z \in \mathcal{S}, \mu_Z = \mu_{Z^c} = \mu_D$, où μ_Z, μ_{Z^c} et μ_D représentent les fonctions moyennes respectivement dans Z, dans Z^c et dans tout le domaine D. L'objectif est donc de tester : $H_0 : \mu_Z = \mu_{Z^c} = \mu_D$ vs. $H_{1,Z} : \mu_Z \neq \mu_{Z^c}$.

Notons $\hat{\mu}_Z$ et $\hat{\mu}_{Z^c}$ les moyennes empiriques dans Z et Z^c, et considérons Γ_{Z,Z^c} l'opérateur de covariance de $\hat{\mu}_Z - \hat{\mu}_{Z^c}$. Pour tester l'égalité des moyennes, une statistique fonctionnelle de Hotelling T^2_{FH} (Joseph et al. 2015, Horváth et al. 2013)est définie comme suit :

$$T^2_{FH} = \sum_{k=1}^{K} \frac{\hat{a}_k^2}{\hat{\lambda}_k},$$

où, $\hat{a}_k = \langle \hat{\mu}_Z - \hat{\mu}_{Z^c}, \hat{\psi}_k \rangle$ représente les scores des composantes principales fonctionnelles pour $k \geq 1$, avec $\hat{\psi}_1, \ldots, \hat{\psi}_k$ et $\hat{\lambda}_1, \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_K$ désignant respectivement les fonctions et les valeurs propres de $\hat{\Gamma}_{Z,Z^c}$: l'estimateur empirique de Γ_{Z,Z^c}. La notation $\langle u, v \rangle = \int_T u(t)v(t)dt$ désigne le produit scalaire dans χ. Le choix de K sera discuté ci-dessous.

Comme dans le cas univarié et multivarié (Kulldorff et al. 2009, Cucala et al. 2017), on suppose que $\Gamma_Z = \Gamma_{Z^c} = \Gamma_X$, ce qui permet d'estimer la covariance de $\hat{\mu}_Z - \hat{\mu}_{Z^c}$ par :

$$\hat{\Gamma}_{Z,Z^c} = \frac{n_Z + n_{Z^c}}{n_Z n_{Z^c}} \hat{\Gamma}_X.$$

où $\hat{\Gamma}_X$ désigne la covariance regroupée, définie par :

$$\hat{\Gamma}_X(\eta) = \frac{1}{n_Z + n_{Z^c} - 2} \Big((n_Z - 1)\hat{\Gamma}_Z(\eta) + (n_{Z^c} - 1)\hat{\Gamma}_{Z^c}(\eta) \Big)$$

pour $\eta \in \chi$. Ici, $\hat{\Gamma}_Z$ et $\hat{\Gamma}_{Z^c}$ représentent les estimateurs empiriques des opérateurs de covariance Γ_Z et Γ_{Z^c}. Comme le soulignent Cucala et al. (2017), Smida et al. (2022), il est crucial de s'assurer que la distribution de l'indice de concentration sous l'hypothèse nulle ne dépend pas de n_Z. Il est important de noter que la statistique de Hotelling T^2_{FH} est équivalente à une version normalisée de la statistique proposée par Horváth et al. (2013). Ainsi, sous H_0, la distribution limite de T^2_{FH} suit une loi du $\chi^2(K)$, qui est fixe et indépendante de n_Z. Ainsi, définissons l'indice de concentration :

$$I(Z) := T^2_{FH},$$

qui permet de comparer des clusters potentiels de tailles variées. Nous introduisons alors la statistique de scan spatial HFSS :

$$\Lambda_{\text{HFSS}} = \max_{Z \in \mathcal{S}} I(Z),$$

et l'agrégat le plus probable :

$$\hat{\text{C}} = \arg\max_{Z \in \mathcal{S}} I(Z),$$

appelé *the most likely cluster* (MLC).

2.2 Calcul de la statistique HFSS

2.2.1 Choix du paramètre seuil K

Le calcul de la statistique de scan HFSS nécessite l'évaluation de l'indice de concentration $I(Z)$ pour chaque cluster potentiel $Z \in \mathcal{S}$, dépendant du paramètre seuil K. Ce paramètre détermine le nombre de valeurs propres et de vecteurs propres de la matrice de covariance

empirique Γ_{Z,Z^c}. Bien que l'indice de concentration soit basé sur la statistique T_H^2, il peut être appliqué avec différents K pour résoudre le problème de test et comparer les résultats. Cependant, lorsque K égale le nombre de mesures k_{max} (fini en pratique), la statistique devient trop grande, souvent menant à un cluster contenant une seule observation. Il est donc conseillé de définir une procédure pour choisir un K approprié, en utilisant par exemple le *cumulative percentage of total variance* (CPV) (Joseph et al. 2015). Cette méthode est standard pour déterminer le nombre de composantes principales à conserver. Pour chaque cluster potentiel Z, la fonction CPV_Z est définie comme suit :

$$CPV_Z(k) = \frac{\sum_{l=1}^{k} \hat{\lambda}_l}{\sum_{l=1}^{k_{max}} \hat{\lambda}_l},$$

où les $\hat{\lambda}_l$ sont les valeurs propres de $\hat{\Gamma}_{Z,Z^c}$. La fonction CPV_Z est croissante et converge vers 1. Elle est calculée pour chaque Z, et en moyennant sur tous les Z, on obtient une fonction $\bar{CPV}$ qui dépend uniquement de k. La prochaine étape consiste à déterminer la valeur optimale de K, ce qui peut être fait de deux manières. Selon Joseph et al. (2015), K peut être choisi là où $\bar{CPV}$ montre une augmentation marginale vers 1. Cette procédure garantit que la même valeur de K est sélectionnée pour tous les clusters potentiels Z, afin que la distribution limite de T_{FH}^2 soit toujours la même quel que soit le cluster Z.

2.3 Significativité du MLC

Une fois la statistique de scan Λ_{HFSS} calculée et le MLC $\hat{C}$ détecté, sa significativité doit être évaluée. Pour cela, nous générons des données par permutation aléatoire des données initiales (*random labelling*). Sur la base de ces permutations, nous considérons $\Lambda_{\text{HFSS}}^{(1)}, \ldots, \Lambda_{\text{HFSS}}^{(T)}$, les statistiques de scan obtenues à partir des données simulées. La p-valeur $\hat{p}$ est alors définie par (Dwass 1957) :

$$\hat{p} = \frac{1 + \sum_{i=1}^{T} \mathbb{1}_{\{\Lambda_{\text{HFSS}}^{(i)} > \Lambda_{\text{HFSS}}\}}}{T+1}.$$

Enfin, le MLC est considéré significatif si la p-valeur associée est inférieure à l'erreur de type I.

3 Applications

3.1 Données simulées

Dans cette section, nous étudions la performance de la statistique de scan HFSS proposée dans la section précédente sur des données simulées. Nous l'avons comparée à (i) la statistique de scan spatial *distribution-free* pour données fonctionnelles (DFSS), (ii) la statistique de scan spatial paramétrique pour données fonctionnelles (PFSS) de Frévent et al. (2021) et (iii) la statistique de scan spatial non paramétrique pour données fonctionnelles (NPFSS) de

Smida et al. (2022). Les localisations spatiales $s_1, \ldots, s_n$ considérées sont les centres des 94 départements français. Le vrai cluster C correspond à un ensemble de départements situés dans la région parisienne. Nous avons testé plusieurs tailles différentes pour C. Pour chaque $i \in \{1, \ldots, 94\}$, les courbes aléatoires X_i appartiennent à $\chi = L^2([0,1], \mathbb{R})$ et sont générées selon la décomposition de Karhunen-Loève. Plusieurs distributions ont été considérées, et trois choix de décalage entre la distribution à l'intérieur et à l'extérieur de C ont été étudiés : linéaire, quadratique et sinusoïdal. Il en ressort que la statistique de balayage HFSS présente les meilleures performances par rapport aux trois autres méthodes. La différence entre les puissances des différents tests est très marquée.

3.2 Données réelles

3.3 Données économiques : taux de chômage en Espagne

Étant donné que l'Espagne a régulièrement compté parmi les pays ayant les taux de chômage les plus élevés de l'Union Européenne, nous avons réalisé une analyse des données sur le chômage. Ces données couvrent chaque trimestre de 2002 à 2022 (un total de 80 trimestres) pour l'ensemble des 47 provinces espagnoles. Nous avons appliqué la statistique de scan HFSS ainsi que les statistiques définies dans la section précédente pour identifier les clusters spatiaux de chômage au niveau provincial.

Résultats : Un cluster significatif, composé de 13 provinces du sud de l'Espagne, a été détecté par les méthodes HFSS, DFFSS et PFSS. Cet agrégat présente des taux de chômage constamment supérieurs à la moyenne nationale. La méthode NPFSS a trouvé un cluster plus large de 17 provinces.

3.4 Données climatiques : différence de températures par rapport aux moyennes en Grande-Bretagne

Nous étudions l'île de Grande-Bretagne, divisée en 145 cellules, pour analyser la différence entre les températures observées de 2010 à 2023 et les moyennes sur la période 1981-2010. Cette différence, calculée comme l'écart avec la température normale, met en évidence un contraste marqué entre le sud chaud et le nord plus frais, avec le sud-est de l'Angleterre présentant les extrêmes les plus élevés. Nous explorons ici si les statistiques de scan peuvent détecter ces tendances.

Résultats : la méthode HFSS a identifié un MLC situé au nord de la Grande-Bretagne, où les différences de température par rapport à la moyenne sont les plus faibles, excepté durant la période 2003-2005, où les températures étaient plus élevées. Ce cluster se distingue des autres méthodes par sa taille et sa localisation, et reflète une zone où les températures sont globalement plus faibles par rapport au reste du pays.

4 Discussion

Dans ce travail, nous avons développé la statistique de scan HFSS pour analyser des données fonctionnelles spatiales, construite à partir d'un test de Hotelling fonctionnel (pour plus de détails, voir Smida et al. 2024). Les résultats de simulation ont montré que la méthode HFSS surpassait les autres en termes de puissance, de taux de vrais positifs et de faux positifs, indépendamment de la taille du cluster, de la distribution ou du type de décalage. Dans nos applications réelles, cette méthode a permis d'identifier des clusters spatiaux significatifs, tant pour les taux de chômage en Espagne que pour les anomalies climatiques en Grande-Bretagne. En particulier, HFSS a détecté des ruptures soudaines dans les courbes, telles que des chocs, ce qui n'était pas toujours le cas avec les autres méthodes de scan fonctionnel.

References

Cucala, L. (2014), 'A distribution-free spatial scan statistic for marked point processes', *Spatial Statistics* **10**, 117–125.

Cucala, L. (2016), 'A Mann–Whitney scan statistic for continuous data', *Communications in Statistics - Theory and Methods* **45**(2), 321–329.

Cucala, L. (2017), Variable window scan statistics: Alternatives to generalized likelihood ratio tests, *in* J. Glaz & M. V. Koutras, eds, 'Handbook of Scan Statistics', Springer New York, New York, NY, pp. 1–16.

Cucala, L., Genin, M., Lanier, C. & Occelli, F. (2017), 'A multivariate Gaussian scan statistic for spatial data', *Spatial Statistics* **21**, 66–74.

Cucala, L., Genin, M., Occelli, F. & Soula, J. (2019), 'A multivariate nonparametric scan statistic for spatial data', *Spatial Statistics* **29**, 1–14.

Dwass, M. (1957), 'Modified randomization tests for nonparametric hypotheses', *The Annals of Mathematical Statistics* **28**(1), 181–187.

Frévent, C., Ahmed, M.-S., Marbac, M. & Genin, M. (2021), 'Detecting spatial clusters in functional data: New scan statistic approaches', *Spatial Statistics* **46**, 100550.

Horváth, L., Kokoszka, P. & Reeder, R. (2013), 'Estimation of the mean of functional time series and a two-sample problem', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **75**(1), 103–122.

Huang, L., Kulldorff, M. & Gregorio, D. (2007), 'A spatial scan statistic for survival data', *Biometrics* **63**, 109–18.

Joseph, E., Galeano San Miguel, P. & Lillo Rodríguez, R. E. (2015), Two-sample Hotelling's T^2 statistics based on the functional Mahalanobis semi-distance, DES - Working Papers. Statistics and Econometrics. WS ws1503, Universidad Carlos III de Madrid.

Kulldorff, M. (1997), 'A spatial scan statistic', *Communications in Statistics - Theory and Methods* **26**, 1481–1496.

Kulldorff, M., Huang, L. & Konty, K. (2009), 'A scan statistic for continuous data based on the normal probability model', *International journal of health geographics* **8**(58).

Kulldorff, M., Mostashari, F., Duczmal, L., Yih, W., Kleinman, K. & Platt, R. (2007), 'Multivariate scan statistics for disease surveillance', *Statistics in medicine* **26**(8), 1824–1833.

Kulldorff, M. & Nagarwalla, N. (1995), 'Spatial disease clusters: Detection and inference', *Statistics in Medicine* **14**(8), 799–810.

Naus, J. (1963), *Clustering of Random Points in Line and Plane*, Harvard University.

Ramsay, J. & Silverman, B. (2005), *Functional Data Analysis*, Springer Series in Statistics, 2nd edn, Springer-Verlag New York.

Smida, Z., Cucala, L., Gannoun, A. & Durif, G. (2022), 'A Wilcoxon-Mann-Whitney spatial scan statistic for functional data', *Computational Statistics & Data Analysis* **167**, 107378.

Smida, Z., Laurent, T. & Cucala, L. (2024), 'A hotelling spatial scan statistic for functional data: application to economic and climate data', `https://theses.hal.science/I3M_UMR5149/hal-04734861v1`. *To appear in Spatial Statistics*.

Régression additive pour réponses Riemaniennes fonctionelles

Jeong Min Jeon [1] & Germain Van Bever [2]

[1] *Ewha Womans University, South Korea, jeongmin.jeon@ewha.ac.kr*
[2] *European Center for Advanced Research in Economics and Statistics (ECARES), Université libre de Bruxelles et Université de Namur, Belgique, germain.van.bever@ulb.be*

Résumé. Cet exposé introduit un modèle de régression additive fonctionnelle dans lequel les réponses sont des fonctions sur une variété Riemanienne générale. Dans ce modèle, les régresseurs sont Euclidiens mais ne sont potentiellement pas observables, de sorte que seule une approximation de ceux-ci est disponibles. Un exemple de tels régresseurs est le cas où ceux-ci sont obtenus par une processus de réduction de dimension d'une variable fonctionnelle. La méthode du "smooth backfitting" est employée pour estimer le modèle additif. Ses propriétés asymptotiques sont étudiées en détail. L'impact des régresseurs imprécis est également quantifié. La méthodologie est illustrée sur des données réelles.

Mots-clés. Analyse fonctionnelle, Données fonctionnelles, Modèle de régression additif, Variété Riemanienne

Abstract. Additive regression is explored for a functional response whose values lie on a general Riemannian manifold. Euclidean predictors that may not be directly observable but are estimable are also addressed, such as component scores obtained from dimension reduction. The smooth backfitting method is employed to estimate the additive model, and its asymptotic properties are derived. Additionally, novel dimension reduction techniques are discussed for general predictors in the presence of the Riemannian functional response. The usefulness of the approach is demonstrated through a real data application.

Keywords. Additive regression model, Functional analysis, Functional data, Riemannian manifol

1 Modèle additif pour variables hilbertiennes

Le traitement de données fonctionnelles ou, plus généralement, de données non-euclidiennes est un sujet largement étudié. Pour traiter ces données, un certain nombre de modèles ont été développés: la régression pour les prédicteurs à valeurs dans une variété (voir par exemple, Jeon *et al.* (2022) et ses références), la régression pour les prédicteurs fonctionnels à valeurs dans une variété ou dont les réponses sont à valeurs multiples (voir Lin *et al.*, 2017), etc. La majorité des travaux dans ce contexte ce concentrent cependant sur la régression nonparamétrique, qui souffre alors du fléau de la dimensionnalité, ou sur la régression paramétrique. A notre connaissance, la régression nonparamétrique pour une

réponse fonctionnelle à valeurs multiples qui évite ce fléau n'a pas été étudiée. Dans cet exposé, nous abordons ce nouveau problème en étudiant la régression additive pour les réponses fonctionnelles à valeurs multiples.

Le modèle classique de régression additive pour la réponse $Y \in \mathbb{R}$ et les régresseurs $(X_1, \ldots, X_d) \in \mathbb{R}^d$ est de la forme

$$Y = \sum_{j=1}^{d} f_j(X_j) + \epsilon,$$

où les $f_j : \mathbb{R} \to \mathbb{R}$ sont les *fonctions composantes* à estimer.

Une méthode pour estimer ces fonctions composantes est le smooth backfitting (SBF, Mammen *et al.*, 1999), connue pour éviter le fléau de la dimensionnalité sous des conditions faibles. Récemment, Jeon et Park (2020) a étudié la méthode SBF pour la régression hilbertienne où la réponse est à valeurs dans un espace de Hilbert général. Cette réponse généralise les réponses euclidiennes, fonctionnelles ou à valeur de densité. Sur la base de cette dernière méthode, Lin *et al.* (2023) a couvert une réponse prenant ses valeurs dans une variété Riemannienne générale en transformant la réponse sur l'espace tangent, qui est également un espace de Hilbert. Lin *et al.* (2023) ne couvre cependant pas le cas des réponses fonctionnelles. En outre, tous les travaux ci-dessus traitent de modèles dans lesquels les fonctions à estimer, telles que les fonctions composantes, sont définies sur $\mathbb{R}$. Récemment, Jeon *et al.* (2022) et Jeon et Van Bever (2024) ont étudié des modèles additifs hilbertiens dont les fonctions composantes sont définies sur des espaces multidimensionnels, de sorte qu'ils couvrent certains effets d'interaction entre prédicteurs scalaires. En outre, Jeon et Van Bever (2024) examine le cas où la réponse et les prédicteurs peuvent ne pas être directement observables mais sont estimables. Il est bien connu que les scores des composantes obtenus par réduction des dimensions, comme l'analyse en composantes principales et l'analyse en composantes principales fonctionnelles, ne sont pas directement observables en raison de structures de moyenne et de covariance inconnues.

Dans cet exposé, nous traitons une réponse fonctionnelle dont le domaine est une variété Riemannienne générale. Pour traiter une telle réponse, nous la transformons dans un espace de Hilbert approprié (espace tenseur Hilbertien) et appliquons un modèle additif hilbertien. La transformation que nous effectuons implique une quantité inconnue, et nous avons donc besoin d'une estimation supplémentaire. Nous couvrons également le cas où les prédicteurs peuvent ne pas être directement observables et où les fonctions composantes sont définies sur des espaces euclidiens multidimensionnels. Nous estimons notre modèle à l'aide de la méthode SBF et dérivons ses propriétés asymptotiques non seulement pour la réponse transformée mais aussi pour la réponse originale. En outre, nous discutons de nouvelles techniques de réduction de la dimension pour les prédicteurs à haute ou infinie dimension associés à la réponse fonctionnelle riemannienne. Dans notre analyse de données réelles, nous abordons le problème de la prévision des trajectoires des typhons à l'aide de prédicteurs euclidiens et à valeurs multiples, qui ne peuvent pas être facilement analysés par les méthodes existantes en raison de la géométrie sphérique de la Terre. Pour traiter les prédicteurs à valeur multiple, nous utilisons une technique de réduction de dimension.

2 Modèle additif pour réponses fonctionnelles Riemaniennes

Soit $\mathcal{T} \subset \mathbb{R}$ un ensemble compact muni d'une mesure de Borel finie ν, et soit $\mathcal{M}$ une variété Riemanienne compacte et connexe. Des exemples de telles $\mathcal{M}$ sont les hypersphères $\mathbb{S}^k = \{p \in \mathbb{R}^{k+1} : \|p\| = 1\}$, les tores $\mathbb{S}^1 \times \cdots \times \mathbb{S}^1$, etc. Soit $W : \mathcal{T} \to \mathcal{M}$ une réponse fonctionnelle.

Pour une fonction mesurable $\mu : \mathcal{T} \to \mathcal{M}$ et $t \in \mathcal{T}$, soit $T_{\mu(t)}\mathcal{M}$ l'espace tangent à $\mathcal{M}$ en $\mu(t)$. Soit $\langle \cdot, \cdot \rangle_{\mu(t)}$ le produit scalaire sur $T_{\mu(t)}\mathcal{M}$, et soit $T\mathcal{M}$ le fibré tangent $\bigcup_{p\in\mathcal{M}} T_p\mathcal{M}$. Soit $\mathfrak{T}(\mu)$ l'ensemble des fonction mesurables $V : \mathcal{T} \to T\mathcal{M}$ satisfaisant $V(t) \in T_{\mu(t)}\mathcal{M}$ pour tout $t \in \mathcal{T}$ et $\int_{\mathcal{T}} \langle V(t), V(t) \rangle \mu(t) d\nu(t) < \infty$. Alors, $\mathfrak{T}(\mu)$ est une espace de Hilbert séparable (voir le Théorème 1 dans Lin et Yao, 2019), appelé l'espace tenseur Hilbertien le long de μ.

Fixons μ_W la moyenne intrinsèque de W sur $\mathcal{M}$. En supposant que, pour tout $\omega \in \Omega$, pour tout $t \in \mathcal{T}$, $W(t) \in \mathcal{M} \setminus Cut(\mu_W(t))$, où $Cut(p)$ est le *cut locus* en $p \in \mathcal{M}$, i.e. l'ensemble des points ayant une image dans la carte en p. Définissons alors la fonction aléatoire $Y : \mathcal{T} \to T\mathcal{M}$ par $Y(t) = Log_{\mu_W(t)}(W(t))$, où Log_p désigne le logarithme Riemanien en p.

Le modèle considéré est le modèle additif Hilbertien (voir par exemple Jeon et Van Bever, 2024)

$$Y = f_0 + \sum_{j=1}^{d} f_j(\xi_j) + \epsilon, \tag{1}$$

où $\xi_j = (\xi_{jl} : 1 \leq l \leq L_j) \in \mathbb{R}^{L_j}$ avec $L_j \in \mathbb{N}$ sont des prédicteurs multivariés, $f_0 \in \mathfrak{T}(\mu_W)$, $f_j : \mathbb{R}^{L_j} \to \mathfrak{T}(\mu_W)$, et $\epsilon \in \mathfrak{T}(\mu_W)$ est un terme d'erreur satisfaisant $\mathbb{E}(\epsilon|\xi) = 0_{\mathfrak{T}(\mu_W)}$ pour $\xi = (\xi_1, \ldots, \xi_d)$. Ici, $0_{\mathfrak{T}(\mu_W)}$ désigne le vecteur nul de $\mathfrak{T}(\mu_W)$, et l'espérance conditionnelle $\mathbb{E}(\epsilon|\xi)$ est définie par intégration de Bochner. Remarquons que le modèle (1) permet l'interaction entre les prédicteurs scalaires dans ξ_j pour chaque $1 \leq j \leq d$.

3 Estimation et théorie asymptotique

L'estimation des fonctions composantes $f_0 \in \mathfrak{T}(\mu_W)$ et f_j s'obtiennent via la méthode de smooth backfitting (voir Mammen et Park, 2006). Ceci requiert cependant l'estimation de μ_W. Un estimateur naturel de la moyenne intrinsèque est donné par

$$\arg\min_{p\in\mathcal{M}} \sum_{i=1}^{n} d^2_{\mathcal{M}}(p, W^i(t)),$$

où $\{W^i : 1 \leq i \leq n\}$ sont les observations de W. Ceci permet alors l'estimation de la moyenne conditionnelle $\hat{\mathbb{E}}(Y|\xi = x) \in \mathfrak{T}(\hat{\mu}_W)$. Remarquons que celle-ci vit dans l'espace tenseur hilbertien associé à la moyenne empirique. Afin de pouvoir introduire le résultat principal du papier associé à cet exposé, nous introduisons le transport parallèle $\Gamma_{\hat{\mu}_W,\mu_W} : \mathfrak{T}(\hat{\mu}_W) \to \mathfrak{T}(\mu_W)$ induit par la connexion de Levi-Civita.

Le résultat principal du papier concerne la théorie asymptotique associé aux estimateurs décrits ci-dessus et donne, sous certaines conditions de régularité des fonctions composantes, les taux de convergence ponctuels, uniformes et L^2 des espérances conditionnelles. Plus précisément, il détaille les taux suivants:

$$\left(\int_S \|\Gamma_{\hat{\mu}_W,\mu_W}(\hat{\mathbb{E}}(Y|\xi = x)) - \mathbb{E}(Y|\xi = x)\|^2_{\mathfrak{T}(\mu_W)} dx \right)^{1/2},$$

$\|\Gamma_{\hat{\mu}_W,\mu_W}(\hat{\mathbb{E}}(Y|\xi = x)) - \mathbb{E}(Y|\xi = x)\|_{\mathfrak{T}(\mu_W)}$, et $\sup_{x\in S} \|\Gamma_{\hat{\mu}_W,\mu_W}(\hat{\mathbb{E}}(Y|\xi = x)) - \mathbb{E}(Y|\xi = x)\|_{\mathfrak{T}(\mu_W)}$.

4 Exemple d'application

Nous détaillerons également l'application suivante, où le modèle additif ci-dessus est utilisé pour prédire la trajectoire de typhons en Asie du Sud-Est. Le package R 'RFPCA' est utilisé pour réduire la dimensions des prédicteurs Riemannniens. Dans cet exemple, $n = 265$ typhons ont été observés entre 2001 et 2021. La Figure 1 ci-dessous représente les trajectoires réelles et estimées des typhons de l'hémisphère Nord.

Bibliographie

Jeon, J. M. and Park, B. U. (2020). Additive regression with Hilbertian responses. *Annals of Statistics*, **48**, 2671-2697.

Jeon, J. M., Park, B. U. and Van Keilegom, I. (2022). Nonparametric regression on Lie groups with measurement errors. *Annals of Statistics*, **50**, 2973-3008.

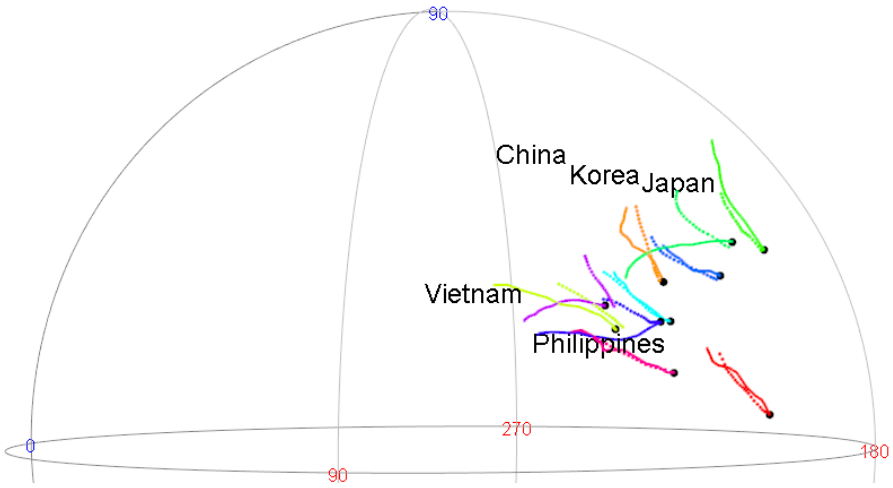


Figure 1: Vraies trajectoires W (ligne pleine) et prédites (lignes pointillées) des typhons décrits ci-dessus. Les points noirs sont les points de départ.

Jeon, J. M. and Van Bever, G. (2024). Multivariate Hilbertian additive regression with general estimated variables. Manuscript.

Lin, L., St. Thomas, B., Zhu, H. and Dunson, D. B. (2017). Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association*, **112**, 1261-1273.

Lin, Z., Müller, H.-G. and Park, B. U. (2023). Additive models for symmetric positive-definite matrices and Lie groups. *Biometrika*, **110**, 361-379.

Lin, Z. and Yao, F. (2019). Intrinsic Riemannian functional data analysis. *Annals of Statistics*, **47**, 3533-3577.

Mammen, E. and Park, B. U. (2006). A simple smooth backfitting method for additive models. *Annals of Statistics*, **34**, 2252-2271.

Données longitudinales

Location-scale linear quantile mixed models for fitting longitudinal data with heteroscedasticity

Angelo Alcaraz [1] & Antoine Barbieri [2]

[1] *Univ. Bretagne Sud, CNRS UMR 6205, LMBA, Vannes, France*
[2] *Univ. Bordeaux, Bordeaux Population Health center, Inserm U1219, Bordeaux, France*

Résumé. La collecte systématique de données longitudinales est très courante dans la pratique, ce qui rend les modèles mixtes très populaires. La plupart des développements autour de ces modèles se limitent à l'ajustement de l'évolution moyenne des mesures répétées sous l'hypothèse d'homoscédasticité. Cependant, le comportement moyen est sensible aux valeurs extrêmes ou aberrantes, ce qui peut fausser les résultats et les interprétations. Par ailleurs, l'hypothèse selon laquelle la variance de l'erreur est constante pour toutes les observations n'est pas toujours vérifiée et l'intérêt de l'inférence peut même être centré sur sa variabilité. Dans ce travail, nous proposons un modèle de régression quantile *location-scale* à effets mixtes permettant de : (1) ajuster un quantile conditionnel de la distribution d'intérêt plutôt que l'espérance conditionnelle, (2) gérer la présence d'hétéroscédasticité dans les données longitudinales, et enfin (3) avoir une meilleure estimation et/ou compréhension de la distribution conditionnelle de la variable d'intérêt. Dans un cadre classique de la régression quantile, nous supposons que la variable d'intérêt suit une loi asymétrique de Laplace où les paramètres de localisation (*location*) et d'échelle (*scale*) sont tous deux définis par un prédicteur linéaire à effets mixtes. Une procédure d'estimation bayésienne est utilisée pour estimer le modèle et est validée par une étude de simulation. Le modèle sera appliqué sur des données réelles pour la modélisation de la pression artérielle au cours du temps.

Mots-clés. Régression quantile, modèle mixte *location-scale*, hétéroscédasticité.

Abstract. The systematic collection of grouped data or repeated measurements is very common in practice, makes mixed models very popular. Most developments around these models are limited on fitting the mean evolution of repeated measurement under the homoscedasticity assumption. However, average behavior is sensitive to extreme values or outlier, which can lead to biased results and interpretations. Furthermore, the assumption that error variance is constant for all observations is not always verified and the inference may even be focused on its variability. In this work, we propose a location-scale linear quantile mixed model to: (1) fit a conditional quantile of the longitudinal outcome distribution instead of conditional expectation, (2) manage the presence of heteroscedasticity in longitudinal data, and finally (3) to have a better estimation and/or understanding of the conditional distribution of the variable of interest. In a classical quantile regression framework, we assume that the variable of interest follows an asymmetric Laplace distribution, where both the location and scale parameters are defined by a linear mixed-effects predictor. A Bayesian estimation procedure is used to estimate the model and is validated by a simulation study. The model will be applied to real data to model blood pressure over time.

Keywords. Quantile regression, location-scale mixed model, heteroscedasticity.

1 Introduction

For the analysis of continuous longitudinal data, or grouped data, it is very common to use a linear mixed model assuming that residual variability is common to all measures (*i.e.* homoscedasticity assumption). However, it may be appropriate for longitudinal data to assume that this variability is group-specific and/or time-dependent. Indeed, heteroscedasticity is a common feature of data and not taking it into account could bias the model's estimation and result in a loss of information on distribution of the response variable. In the literature, location-scale mixed models have been proposed to overcome this assumption (Hedeker et al., 2008; Courcoul et al., 2024). As the name suggests, parametric model fitting concerns both the location parameter and the scale parameter of the distribution of interest. Under the Gaussian assumption, this means that both expectation (*i.e.* location parameter) and residual variance (*i.e.* scale parameter) are defined by a linear predictor of their own. Therefore, residual variability is fitted as a function of covariates and group-specific random effects.

In addition to the presence of heteroscedasticity, data can present asymmetric distribution, extreme values and outliers which can be affect the estimation of distribution parameters. The traditional Gaussian approach of relating the conditional mean of the response to the explanatory variables is therefore unsuitable. An alternative to the mean for assessing the central behavior of the distribution is to study the median, which is more robust to the presence of outliers. For this, quantile regression would seem to be the ideal solution. In fact, it can be applied to any quantile of the response variable distribution, and can handle the presence of asymmetry in the data.

In this work, we propose a location-scale linear quantile mixed model. The aim is to take into account the presence of heteroscedasticity and asymmetry in longitudinal data, while being robust to extreme values. To achieve this, we consider a parametric framework based on the asymmetric Laplace ($\mathcal{AL}$) distribution classically used in quantile regression (Geraci and Bottai, 2007). The location and scale parameters of this distribution are both adjusted by linear mixed predictors, while the skewness parameter is fixed according to the quantile of interest or the asymmetry present in the data.

2 The location-scale linear quantile mixed model

In a classical linear mixed model, the response variable is assumed to be normally distributed, and the focus is on modeling the location parameter (*i.e.* the expectation) through a linear predictor. In quantile regression, it is natural to assume that the response variable is distributed according to an $\mathcal{AL}$ distribution (Geraci and Bottai, 2007). The latter is defined by a location parameter $\mu \in \mathbb{R}$ and a scale parameter $\sigma \in \mathbb{R}^+$ as for the normal distribution, but also by a skewness parameter $\tau \in]0,1[$. Since the location parameter μ corresponds to the quantile of order τ of the distribution, classical quantile linear regression consists in defining the location parameter by a linear predictor. In this work, we also adjust the scale parameter by a linear predictor to take into account that residual variability is not common between observations, groups or over time for longitudinal data.

Let $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{in_i})^\top$ denotes the vector of the n_i measurements for the group $i (i = 1, \ldots, n)$. Tthe response variable conditionnally to random effects is assumed to be distributed from the $\mathcal{AL}$ distribution such as

$$Y_{ij}|\mathbf{b}_i, \mathbf{u}_i \ \sim \ \mathcal{AL}\left(\ \mu_{ij,\tau},\ \sigma_{ij,\tau},\ \tau\ \right), \tag{1}$$

where $\tau \in]0, 1[$ is the skewness parameter, and $\mu_{ij,\tau}$ and $\sigma_{ij,\tau}$ are the location and scale parameters respectively defined by

$$\mu_{ij,\tau} \ = \ \mathbf{x}_{ij,\mu}^\top \boldsymbol{\beta}_\tau \ + \ \mathbf{z}_{ij,\mu}^\top \mathbf{b}_{i,\tau} \tag{2}$$

$$\sigma_{ij,\tau} \ = \ exp\left(\mathbf{x}_{ij,\sigma}^\top \boldsymbol{\xi}_\tau + \mathbf{z}_{ij,\sigma}^\top \mathbf{u}_{i,\tau}\right) \tag{3}$$

with $\boldsymbol{\beta}_\tau$ is the p_β-length vector of location fixed effects associated with design vector $\mathbf{x}_{ij,\mu}$, $\mathbf{b}_{i,\tau} \sim \mathcal{N}(\mathbf{0}, \Sigma_b)$ is a grouped-specific random effect vector associated with design vector $z_{ij,\mu}$, $\boldsymbol{\xi}_\tau$ is the p_σ-length vector of scale fixed effects associated with design vector $\mathbf{x}_{ij,\sigma}$, $\mathbf{u}_{i,\tau} \sim \mathcal{N}(\mathbf{0}, \Sigma_u)$ is the grouped-specific random effect vector associated with design vector $\mathbf{z}_{ij,\sigma}$. The random effect vectors $b_{i,\tau}$ and $u_{i,\tau}$ are assumed to be independent, as is classically done linear regression.

In the quantile regression literature, the parameter τ is most often fixed as it defines the order of the quantile sought (Koenker, 2005). However, when we are interested in the formulation of a quantile regression problem using the $\mathcal{AL}$ distribution, this parameter will influence the response distribution. Consequently, it is also possible to choose τ in such a way that the $\mathcal{AL}$ distribution chosen shares certain properties of the observed distributions of response response.

3 Estimation procedure

The estimation procedure considers a Bayesian framework and a rewriting of $\mathcal{AL}$ distribution. Instead of directly considering the $\mathcal{AL}$ distribution, an alternative is to use its Gaussian mixture representation for quantile regression model estimation (Kozumi and Kobayashi, 2011; Kotz et al., 2012). For $Y \sim \mathcal{AL}(\mu, \sigma, \tau)$, the probability density can be defined by

$$f_Y(y) = \int_{\mathbb{R}^+} f_{Y|W}(y|W = w)\, f_W(w)\, dw \tag{4}$$

with

$$\begin{cases} Y|\mu, \sigma, \tau, W = w \ \sim \ \mathcal{N}\Big(\mu + c_1(\tau)\, w,\ c_2(\tau)\, \sigma w\Big) \\ W|\sigma \ \sim \ \mathcal{E}xp\left(\frac{1}{\sigma}\right) \end{cases},$$

with $c_1(\tau) = \frac{1-2\tau}{\tau(1-\tau)}$, $c_2(\tau) = \frac{2}{\tau(1-\tau)}$ and W an auxiliary random variable exponentially distributed such as $\mathbb{E}(W) = \sigma$. Therefore, we define our model as a hierarchical Bayesian

model with

$$\begin{aligned} Y_{ij}|W_{ij}=w_{ij},\mathbf{b}_{i,\tau},\mathbf{u}_{i,\tau},\boldsymbol{\theta}_\tau &\sim \mathcal{N}\Big(\mu_{ij,\tau}+c_1(\tau)\,w_{ij},\ c_2(\tau)\,\sigma_{ij,\tau}w_{ij}\Big) \\ W_{ij}|\boldsymbol{\xi}_\tau,\mathbf{u}_{i,\tau} &\sim \mathcal{E}xp\left(\frac{1}{\sigma_{ij,\tau}}\right) \quad \text{with } \mathbb{E}(W_{ij})=\sigma_{ij,\tau}\ , \\ \mathbf{b}_{i,\tau}|\Sigma_b &\sim \mathcal{N}\Big(\mathbf{0},\ \Sigma_b\Big) \\ \mathbf{u}_{i,\tau}|\Sigma_u &\sim \mathcal{N}\Big(\mathbf{0},\ \Sigma_u\Big) \end{aligned} \tag{5}$$

where $\boldsymbol{\theta}_\tau=(\boldsymbol{\beta}_\tau,\Sigma_b,\boldsymbol{\xi}_\tau,\Sigma_u)^\top$ is the vector of model parameters.

In Bayesian framwork, the interest focuses on posterior distribution of parameters:

$$\pi(\boldsymbol{\theta}_\tau,\boldsymbol{b},\boldsymbol{u}|\boldsymbol{y},\boldsymbol{t},\boldsymbol{\delta})\propto\prod_{i=1}^{n}\prod_{j=1}^{n_i} f(y_{ij}|w_{ij},\mathbf{b}_{i,\tau},\mathbf{u}_{i,\tau},\boldsymbol{\theta}_\tau)\,f(w_{ij}|\boldsymbol{\theta}_\tau)\,\phi(\mathbf{b}_{i,\tau}|\Sigma_b)\,\phi(\mathbf{u}_{i,\tau}|\Sigma_u)\,\pi(\boldsymbol{\theta}_\tau)$$

where $f(.)$ are density functions associated with distributions defined in (5), ϕ denotes the zero-mean normal density function for subject-specific random effects, and $\pi(\boldsymbol{\theta}_\tau)$ is the prior distribution of parameters which is assumed to be independent. Concerning prior distributions of parameters, we considered normal distribution for regression parameters and an inverse Wishart ($\mathcal{IW}$) distribution for the covariance matrices of random effects:

$$\boldsymbol{\beta}_\tau\sim\mathcal{N}_{p_\beta}\Big(\boldsymbol{\mu}_\beta,\ \Sigma_\beta\Big),\quad \boldsymbol{\xi}_\tau\sim\mathcal{N}_{p_\sigma}\Big(\boldsymbol{\mu}_\xi,\ \Sigma_\xi\Big),\quad \Sigma_b\sim\mathcal{IW}\Big(\mathbf{a}_b,\ A_b\Big)\text{ and }\Sigma_u\sim\mathcal{IW}\Big(\mathbf{a}_u,\ A_u\Big).$$

These priors can be considered as vague or data driven using previous estimated model without random effects.

Posterior parameter samples are obtained using the R-package `BeQut`. It has been developed for the estimation of the model using the `JAGS` software. The development version of `BeQut` is available on Github at the following link: `https://github.com/AntoineBbi/BeQut`.

4 Results

The estimation procedure has been validated through a simulation study. This study highlights an unbiased estimation of the parameters, a good coverage rate of the parameters, and good convergence of the MCMC chains. The application of the model will be illustrated for the longitudinal modeling of patients' systolic blood pressure.

5 Discussion

In this work, we propose a location-scale linear quantile mixed model taking into account the presence of heteroscedasticity and asymmetry in longitudinal data, while being robust to extreme values. Based on the asymmetric Laplace distribution, our model can assess the

impact of covariates on both the individual evolution of the location parameter and the scale parameter. Then, the choice of the skewness parameter allows to manage the asymmetry in the data and/or to select the order of the quantile of interest. Finally, since this distribution is fully characterized by its parameters, the interest can extend to the study of the complete conditional response distribution (Kneib et al. 2023).

References

Courcoul, L., Tzourio, C., Woodward, M., Barbieri, A., and Jacqmin-Gadda, H. (2024). A location-scale joint model for studying the link between the time-dependent subject-specific variability of blood pressure and competing events. *arXiv* :2306.16785.

Geraci, M and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. Biostatistics, 8(1):140-154.

Hedeker, D., Mermelstein, R. J., and Demirtas, H. (2008). An Application of a Mixed-Effects Location Scale Model for Analysis of Ecological Momentary Assessment (EMA) Data. *Biometrics*, 64(2) :627–634.

Roger Koenker. (2005). *Quantile Regression*. Cambridge University Press.

Kneib T., Silbersdorff, A., Säfken, B. (2023). Rage Against the Mean – A Review of Distributional Regression Approaches. *Econometrics and Statistics*, (26):99-123.

Kotz, S., Kozubowski, T., and Podgorski, K. (2012). The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance. Springer Science & Business Media.

Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation and Simulation*, 81(11):1565-1578.

Random forests using longitudinal predictors

Justine Remiat[1], Cécile Proust-Lima[2], Robin Genuer[3]

[1] *Univ. Bordeaux, INSERM, INRIA, BPH, U1219, France, justine.remiat@u-bordeaux.fr*
[2] *Univ. Bordeaux, INSERM, BPH, U1219, France, cecile.proust-lima@u-bordeaux.fr*
[3] *Univ. Bordeaux, INSERM, INRIA, BPH, U1219, France, robin.genuer@u-bordeaux.fr*

Résumé. Les forêts aléatoires (Breiman, 2001) sont un outil de prédiction efficaces, notamment dans le contexte de grande dimension. Toutefois, elles ne sont pas adaptées aux données longitudinales fréquemment collectées dans les études de santé. Pour remédier à cela, Capitaine et al. (2024) ont introduit les forêts aléatoires de Fréchet, qui peuvent traiter tout type de données dans un espace métrique, en utilisant une distance adaptée à chaque type de donnée (image, trajectoire, etc.).
L'objectif de cette étude est d'implémenter les forêts aléatoires de Fréchet utilisant des trajectoires en exploitant pleinement la flexibilité de la distance discrète généralisée de Fréchet et d'évaluer leur performance dans la prédiction d'une variable continue à partir de variables longitudinales.

Dans le cas des trajectoires, la distance discrète généralisée de Fréchet est utilisée. Cette distance dépend d'un paramètre de décalage temporel, nommé timescale, modifiant le comportement de la distance. Nous avons proposé deux implémentations : une où le timescale est définie comme hyper-paramètre par l'utilisateur, et une autre où elle est tirée aléatoirement à chaque nœud de chaque arbre pour explorer tous les types d'association possibles. Une étude de simulation a été réalisée pour illustrer la flexibilité des forêts aléatoires de Fréchet à capturer différents types d'associations : (i) sensible à la temporalité, (ii) sensible à la forme globale de la trajectoire, et (iii) combinaison des deux. Cette méthode a également été appliquée pour prédire la probabilité de développer une démence à partir de trajectoires de santé physique et cérébrale dans une cohorte en population (étude 3C).

Pour les données simulées, nous avons observé que le timescale influence fortement les performances des forêts. Un timescale élevé améliore les performances de la forêt lorsque la variable à prédire est associée à des paramètres temporels (i) tandis que lorsque la variable à prédire est liée à des paramètres de forme globale des prédicteurs (ii), un timescale faible semble plus judicieux. Comparées à des forêts classiques utilisant des résumés des trajectoires, les forêts aléatoires de Fréchet ont systématiquement montré de meilleures performances (MSE) dans les trois scénarios, bien que les deux types de forêts aient extrait les marqueurs les plus pertinents.
Sur les données réelles, les forêts aléatoires de Fréchet ont montré une erreur de prédiction plus faible que les forêts classiques malgré des données plus irrégulières et moins denses, tout en identifiant des marqueurs prédictifs de démence cohérents avec la littérature.

Grâce au paramètre de timescale permettant de moduler le comportement de la distance de Fréchet, les forêts aléatoires de Fréchet sont un outil flexible de prédiction à partir de données longitudinales.

Mots-clés. Distance, données longitudinales, forêts aléatoires de Fréchet, prédiction

Abstract. Random Forests (Breiman, 2001) are an effective predictive tool, particularly in high-dimensional settings. However, they are not well-suited for longitudinal data collected over time. To address this limitation, Fréchet Random Forests (Capitaine et al. 2020) were proposed. They can handle any type of data within a metric space by using a distance tailored to each data type (e.g., images, trajectories). This work aimed to implement the Fréchet Random Forest for trajectory data, fully exploiting the flexibility of the generalized discrete Fréchet distance; and evaluate the performance of the Fréchet Random Forest in predicting a continuous outcome using longitudinal inputs.

The Generalized Discrete Fréchet Distance depends on a time-shifting parameter, called timescale, which modifies its behavior. We proposed two implementations: the time-scale defined as an hyper parameter or the time-scale randomly drawn at each tree node to explore all time sensitivity behaviors. A simulation study has been conducted to illustrate the flexibility of the Fréchet Random Forest to capture different scenarios of association: (i) time-sensitive association (ii) shape-sensitive association and (iii) a mix of both. We then apply the method to data from a population-based cohort to predict the risk of dementia from clinical marker trajectories.

The simulations illustrated the flexibility of the Fréchet Random Forests to adapt to different types of associations with the timescale tuning. The Fréchet Random Forests also demonstrated better predictive performance (MSE) across all three scenarios compared to classical Random Forests with pre-determined features.
On the application data, the Fréchet forests outperformed classical forests, even with more irregular and sparse data, while similarly identifying predictive markers.

Thanks to its tunable timescale parameter that can adapt to different structures of association, the Fréchet Random Forest constitutes a flexible tool for prediction based on longitudinal data.

Keywords. distance, longitudinal data, Fréchet random forest, prediction

1 Introduction

In Public Health, data are becoming increasingly complex, with high-dimensional datasets made of variables often collected at multiple time points. Studies are also becoming more ambitious, aiming to predict outcomes or future trends of biomarkers from the history of a large set of repeated predictors.
Random Forests (RF) (Breiman, 2001) are among the state-of-the-art machine learning methods for prediction, especially in in high-dimensional settings . However, they are not inherently suited for the longitudinal data collected in cohort studies and clinical trials.
To address this limitation, several extensions of Random Forests have been proposed. Notably, Fréchet Random Forests (Capitaine et al. 2024) constitutes a promising framework. It is a generalization capable of handling any type of input or output data, provided they

belong to a metric space.
The objective of this study was to (1) implement Fréchet Random Forests (FRF) fully exploiting the flexibility of the generalized discrete Fréchet distance (Genolini et al., 2016); (2) evaluate the performance of FRF in predicting a continuous outcome from longitudinal inputs under various association mechanisms, and (3) illustrate its use to predict dementia risk from trajectories of various predictors in a population-based cohort.

2 Fréchet Random Forest (FRF)

CART trees were introduced by Breiman et al. (1984) as a method for building decision trees using a learning sample $\mathcal{L}_q = \{(X_1, Y_1), \ldots, (X_q, Y_q)\}$ where the variables $X \in \mathbb{R}^p$ are predictors and $Y \in \mathbb{R}$ is the response variable (for regression trees). A CART tree is constructed through recursive binary splits of $\mathbb{R}^p$. Each split is chosen as the one that maximizes the reduction in heterogeneity between the resulting child nodes

Random Forests, introduced by Breiman (2001), are an ensemble learning method based on decision trees. The approach involves creating bootstrap samples of the learning set, with each sample used to construct an individual randomized tree, known as a Random Input tree. In random input tree, at each split, only a random subset of the explanatory variables is used to optimize the split of this node into two child nodes. The number of candidate variables for each split is denoted by m_{try} .

To make predictions, the outputs of all the trees in the forest are aggregated. For regression, this is done by averaging the predictions of the individual trees.

A key performance metric for RF is the Out-Of-Bag (OOB) error. For each tree, OOB observations are those not included in its associated bootstrap sample. Once the forest is built, for each observation, an OOB prediction is obtained by aggregating predictions from all trees for which this observation was OOB, and, in a regression setting, the Mean Squared Error (MSE) is then computed from these predictions..

Random Forests also include a feature for measuring variable importance, which quantifies the contribution of each predictor variable to the model's performance.

Fréchet Random Forests (Capitaine et al., 2024), (implemented in the `FrechForest R package`), generalize RF to handle complex inputs like curves or images.
Fréchet random trees, that we use, follow the same process as Extremely Randomized Trees (ERT) (Geurts et al.,2006) in which the split is either fully or partially selected at random, and the splitting rule is adapted for different input types. For curve inputs, two curves are randomly chosen as representatives for child nodes, and observations are assigned to the left or the right child node based on their distances to those representatives. This process is repeated `ntry` times and the split that maximizes the decrease in node heterogeneity is selected. The same process is performed for continuous inputs, whereas for categorical inputs, binary partitions of the set of categories are randomly selected directly from all possible partitions.. The assignment metric for curves is the Generalized Discrete Fréchet Distance (GDFD) as Euclidean distance is unsuitable for curves with varying point counts.
The Fréchet distance was introduced by Maurice Fréchet (1906). The formula for the Fréchet distance between A and B (two curves from $[0, t_{\max}]$ to $\mathbb{R}$) is as follows:

$F(A,B) = \inf_{\alpha,\beta} \max_t \{d(A(\alpha(t)), B(\beta(t)))\}$, where d is the Euclidean distance, and α and β are increasing re-parametrizations.

The GDFD comes with a parameter λ (the timescale parameter) that controls the behavior of the distance regarding its sensitivity to time. The discrete Fréchet distance is computed after an affine transformation of A and B: $\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} \lambda \times x \\ y \end{pmatrix}$ with $\lambda \in \mathbb{R}_+^*$.

When $\lambda \to 0$, GDFD resembles Dynamic Time Warping (DTW), where time shifts have no effect;
and when $\lambda \to +\infty$, GDFD behaves like the max Euclidean distance, where horizontal shifts significantly impact distances (Genolini et al., 2016).

3 Simulations study

The simulation study aimed to assess the performances of the FRF to predict a continuous outcome from various marker trajectories, and compare them with those obtained by classical RF (implemented, e.g., in the `ranger` R package) using extracted features. Simulations included 500 individuals with 14 longitudinal markers measured across 16 visits (fig.1). Markers were generated according to mechanistic mixed models, change-point mixed models and spline-trajectory mixed models. Three scenarios linked the outcome (Y) to parameters associated with two longitudinal markers:

i) parameters linked to the temporal trend of markers (time-sensitive);
ii) parameters linked to the global shape of markers;
iii) parameters linked to the shape and timing of trend of markers.

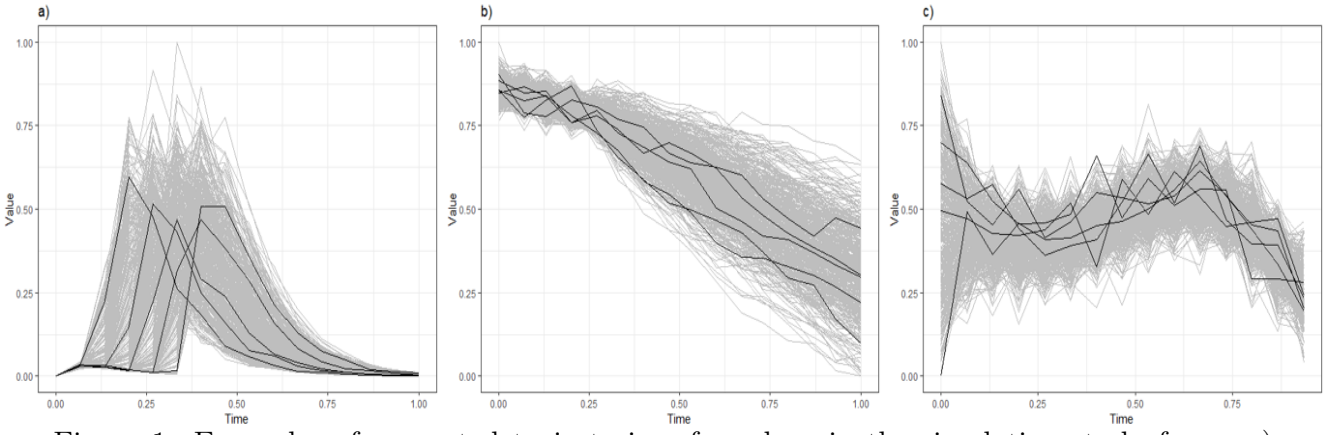


Figure 1: Examples of generated trajectories of markers in the simulation study from: a) a mechanistic model, b) a change-point mixed model, c) mixed model with a nonlinear trajectory approximated by splines.

As a comparison, two types of classical RF were built:

- **RF**: Using the minimum, maximum, mean, and standard deviation of each marker (56 input variables). For tuning, we considered $m_{\text{try}} \in [\![1; 56]\!]$.

- **RF++**: Using the minimum, maximum, mean, standard deviation, time of minimum, and time of maximum for each marker (84 input variables). For tuning, we considered $m_{\text{try}} \in [\![1; 84]\!]$.

One dataset was simulated to tune the hyperparameters, and 500 datasets were simulated to evaluate performance. Performance was measured using out-of-bag (OOB) mean squared error (MSE).

Variable importance was also computed across the 500 datasets to assess whether each type of forest correctly identified the relevant markers as important.

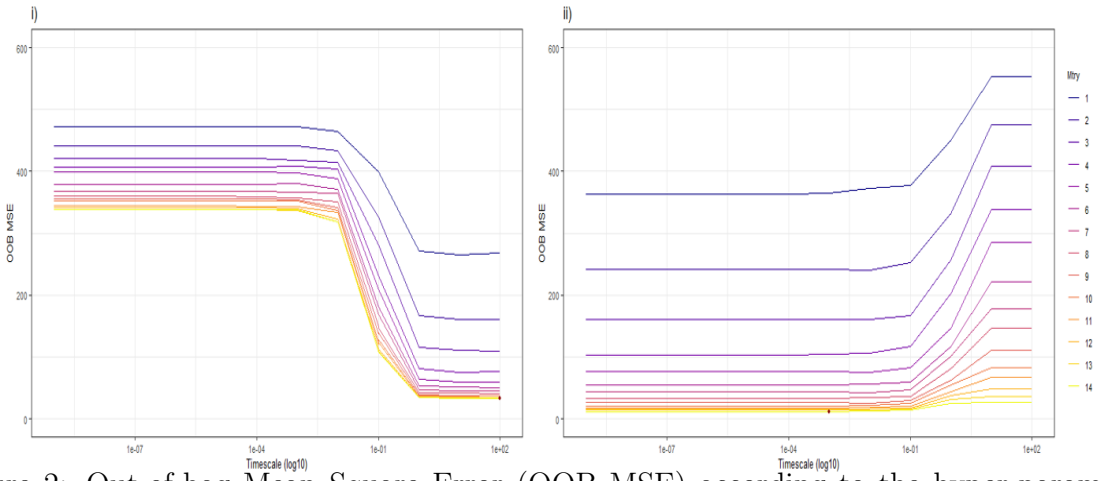


Figure 2: Out-of-bag Mean Square Error (OOB MSE) according to the hyper-parameters of the Fréchet Random Forest: timescaling and number of trajectories tried (mtry) in two scenarios:
i) time-sensitive scenario , and ii) time-insensitive scenario .

Tuning focused on $mtry \in [1; 14]$ and $\lambda \in [10^{-9}; 10^{2}]$. Higher λ were selected for the time-sensitive scenario (i) while lower λ performed better for global shape scenario (ii) (fig.2). Across all scenarios, FRF outperformed classical RF in MSE, and both methods correctly retrieved the associated markers.

4 Application

We applied the method to predict the probability of developing dementia using the trajectories of 28 markers of cerebral and physical health in the population-based French cohort study 3C. We considered the 817 participants from Bordeaux city who were free of dementia at their 10-year follow-up visit, and used 22 markers recorded during the first 10 years of follow-up, such as cognitive scores, health indicators, and 6 sociodemographic variables measured at baseline. Two dementia outcomes were considered at an horizon of 5 years: the pseudo-value of the probability to develop dementia, and the restricted mean survival time. The FRF continued to outperform classical RF, despite the repeated data being sparser and more irregular than the simulated data. Overall both methods identified the same key predictors of dementia risk.

5 Conclusion

FRF demonstrate strong performance, making them a promising tool for handling complex longitudinal data. The flexibility of the method thanks to the tunable timescale parameter, helps capturing different structures of association. This adaptability, combined with robust results, highlights their potential for various applications in predictive modeling.

Bibliography

Balelli I, Pasin C, Prague M, Crauste F, Effelterre TV, Bockstal V, et al. A model for establishment, maintenance and reactivation of the immune response after vaccination against Ebola virus. *Journal of Theoretical Biology*. 2020 Jun;495:110254.

Breiman L. Random Forests. *Machine Learning*. 2001 Oct 1;45(1):5–32.

Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and Regression Trees. *New York: Chapman and Hall/CRC*; 2017. 368 p.

Capitaine L., Bigot J., Thiébaut R., Genuer R., Fréchet random forests for metric space valued regression with non euclidean predictors, *Journal of Machine Learning Research*, 25(355):1-41 (2024)

Fréchet MM. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884–1940)*. 1906; 22(1):1–72.

Genolini C, Ecochard R, Benghezal M, Driss T, Andrieu S, Subtil F. kmlShape: An Efficient Method to Cluster Longitudinal Data (Time-Series) According to Their Shapes. *PLoS One*. 2016 Jun 3;11(6):e0150738.

Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Machine Learning*. 2006 Apr;63(1):3–42.

Time-penalized trees: consistency results and simulations

Mathias Valla [1,2,†] & Xavier Milhaud [1,2,‡]

[1] *Chaire DIALog, Fondation du risque, Institut Louis Bachelier, Paris, France*
[2] *I2M, Aix-Marseille Université, Marseille, France*
[†] *Email auteur: mathias.valla@gmail.com*
[‡] *Email auteur: xaviermilhaudresearch@gmail.com*

Résumé. Cet article étend l'étude des arbres pénalisés par le temps (TpT), un nouvel algorithme basé sur les arbres de décision et conçu pour les données longitudinales dont les variables évoluent dans le temps. TpT introduit une pénalité temporelle exponentielle dans le critère de division d'un arbre de décision classique, ce qui permet un partitionnement conjoint de l'espace des covariables et du temps. Nous présentons une analyse théorique montrant que, sous des hypothèses raisonnables concernant la durée d'observation des sujets et la distribution des gains d'homogénéité obtenus à chaque noeud, un TpT peut être un estimateur consistent. Plus précisément, nous prouvons que (i) lorsque la taille de l'échantillon global augmente, le nombre d'observations dans chaque sous-arbre temporel tend vers l'infini, et (ii) ces sous-arbres peuvent atteindre une profondeur illimitée. Une preuve de consitance de TpT s'appuie ensuite sur les résultats classiques de consistence des arbres de décision classiques. Nous complétons ces résultats par des simulations qui illustrent comment, pour des distributions de durée d'observation à support fini ou sous-exponentielles et si les gains d'homogénéité suivent une distribution raisonnable, des sous-arbres temporels convenablement choisis continuent à se diviser indéfiniment. Bien qu'une comparaison expérimentale approfondie avec les modèles longitudinaux existants et une méthode de pénalisation plus générale restent des pistes de travail pour l'avenir, nos résultats soulignent le potentiel de TpT en tant que méthode interprétable et efficace pour la prise de décision dynamique dans des contextes longitudinaux.

Mots-clés. arbre de décision, covariables variables dans le temps, données longitudinales, algorithme, partitionnement récursif, cohérence, interprétabilité

Abstract. This paper extends the study of Time-penalized Trees (TpT), a novel tree-based algorithm designed for longitudinal data with time-varying covariates and responses. TpT introduces an exponential time-penalty into the splitting criterion, enabling joint partitioning of covariate space and time. We provide a theoretical analysis showing that, under mild assumptions on subject durations and the distribution of splitting gains, TpTs can achieve consistency. Specifically, we prove that (i) as the global sample size tends to infinity, the number of observations in each time-based subtree also tends to infinity, and (ii) these subtrees can grow unboundedly in depth, leveraging classical decision trees consistency results. We complement these findings with a simulation study illustrating how, for finite-support or subexponential duration distributions, the number of subjects reaching any given time node increases without bound, while suitably chosen time subtrees continue to split indefinitely if impurity gains follow a distribution with a nontrivial supremum. Although a thorough experimental comparison with existing longitudinal models, as well as a more flexible penalty framework remain avenues for future work, our results highlight the potential of TpT as an interpretable and computationally efficient method for dynamic decision making in data-intensive longitudinal settings.

Keywords. decision tree, time-varying covariates, longitudinal data, algorithm, recursive partitioning, consistency, interpretability

1 Motivation

Decision trees are widely used in machine learning due to their **interpretability** and simplicity. By recursively splitting the feature space, each node serves as a rule, making it easy for analysts to follow the decision path. This rule-based nature also facilitates **visualization** through flowchart-like diagrams, which enhance transparency in fields such as financial risk management, medical diagnostics, and regulatory compliance.

While ensemble methods like Random Forests and Gradient Boosted Trees have been extensively studied, the theoretical properties of a *single* decision tree have received less attention. Deep single trees can be unstable and exhibit high variance; however, Klusowski Klusowski and Tian (2024) recently demonstrated that under natural sparsity conditions, even single decision trees can be consistent. This insight provides a counterpoint to the prevailing emphasis on ensemble consistency (e.g., Biau et al. (2008); Biau (2012); Biau and Scornet (2016) and Scornet et al. (2015); Scornet (2016)). Empirical evidence further supports that tree-based models often perform competitively on tabular data—sometimes outperforming deep neural networks, especially on moderate-sized datasets Grinsztajn et al. (2022).

Standard decision trees assume each sample is an independent observation, a premise that becomes problematic in *longitudinal* settings where subjects are measured repeatedly over time. A common remedy is to split each subject's time series into "pseudo-individuals" (see Fu and Simonoff (2016); Yao et al. (2020)), effectively treating each time point as independent. This, however, disregards the inherent correlation between observations and fragments each subject's unique trajectory, potentially placing data from the same individual in multiple branches.

Motivated by the need for consistent and interpretable modeling of longitudinal data, Valla (2024) introduced Time-penalized Trees (TpT). In TpT, time is treated as a fundamental splitting axis alongside other covariates but is penalized exponentially to discourage splits that occur late or far apart in time. This design allows each subject to trace a *single path* from the root to a terminal node, clearly reflecting their temporal progression. Enhanced visualizations can also be produced by annotating time-based splits, thereby emphasizing the temporal context in which certain rules become operative.

Here, we investigate the theoretical foundations and practical advantages of TpT. We describe its construction, provide theoretical justification for the exponential time penalty using memoryless and maximum entropy arguments, extend consistency results to the longitudinal framework, and illustrate key steps of our proof with simulation studies.

2 What is a Time-penalized Tree?

Time-penalized Trees (TpT) are tailored for longitudinal data by splitting on both covariates and time, unlike standard CART-like methods (see Breiman et al. (1984) or Quinlan (1993)) that split only on covariates. TpT partitions the joint space $\mathcal{X} \times \mathcal{T}$ and uses a time penalty to uncover time-adaptive structures while keeping the tree interpretable. This approach effectively captures evolving feature/outcome relationships, making it well-suited for repeated measurements.

2.1 Building process

In a Time-penalised Tree and in full generality, a single split of a parent node $g^p_{t_p}$ results into three nodes: the left child node $g^l_{t_c}$, the right child node $g^r_{t_c}$, and the duration leaf $g^d_{t_p}$, as illustrated in Figure **??**. To obtain such a split, we have to maximize a time-penalised split criterion, as

$$G_\gamma(g^p_{t_p}) = max\left(G(g^p_{t_p}) \cdot e^{-\gamma\cdot(t_c - t_p)}; 0\right), \tag{1}$$

with $\gamma \in \mathbb{R}^+$ and

$$G(g^p_{t_p}) = \left[I(g^p_{t_p}) - \left(\frac{\mathcal{N}(g^l_{t_c})}{\mathcal{N}(g^p_{t_p})} I(g^l_{t_c}) + \frac{\mathcal{N}(g^r_{t_c})}{\mathcal{N}(g^p_{t_p})} I(g^r_{t_c}) + \frac{\mathcal{N}(g^d_{t_p})}{\mathcal{N}(g^p_{t_p})} I(g^d_{t_p})\right)\right].$$

The function $I(.)$ is an impurity function, typically the within-node sample variance of the response (or mean squared error) in the regression setting or the Gini or Entropy measures for classification.

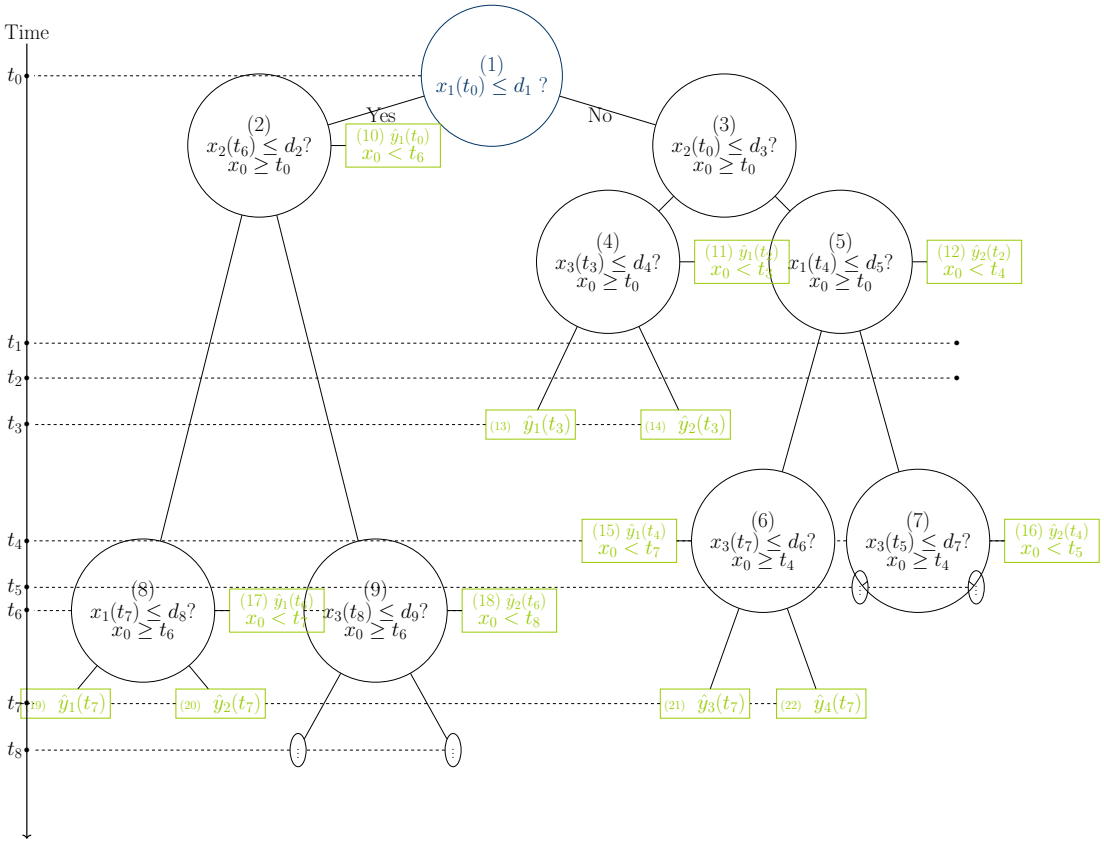


Figure 1: Illustration of a `TpT`

A convenient way to visualize a TpT is provided in Figure 1, where we added some node identifiers for ease of description. In the latter illustration, all partitions of the initial space $\mathcal{X} \times \mathcal{T}$ are represented. Nodes appears as circles, whereas leaves are green rectangles. The time scale, on the left-hand side, allows to highlight that the time dimension plays a key role for the explainability of some covariates on the response, an important aspect in longitudinal studies.

In this representation, it is clear that the union of all *t-subtrees* enables to recover the entire initial product space $\mathcal{X} \times \mathcal{T}$, and that each *t-subtree* forms a partition of this space.

3 Consistency of a TpT

We detail a four-step approach to proving consistency for the overall TpT estimator. Each step addresses a separate challenge posed by the time dimension and the recursive structure of TpT.

Step 1: Prove that a TpT is a sequence of CART-like trees.

Step 2: Show that $\forall t,\ n \to +\infty \ \Rightarrow\ n(t) \to +\infty$**.**

Step 3: For a given t**-subtree,** $n(t) \to +\infty \ \Rightarrow\ K(t) \to +\infty$**.**

Step 4: Apply Klusowski–Tian consistency results to each t**-subtree.** Once each t-subtree is shown to possess the ability to grow unboundedly in depth and sample size, existing consistency theorems for single decision trees become applicable on a subtree-by-subtree basis. By combining consistency across all ts where it can be achieved, we obtain a global consistency guarantee for the full TpT:

Theorem 1 (Consistency of Time-penalized Trees) *Let $\mathcal{D}_{long}$ be a longitudinal dataset of size n, in which each subject i has a duration variable $X_0^{(i)}$ with survival function S, and let $\gamma \geq 0$ be the time-penalty parameter used in the TpT splitting criterion. Suppose:*

(A) *The distribution of X_0 satisfies one of the following:*

- *X_0 has finite support $[\,0,\ T_{\max}\,]$, or*
- *X_0 is supported on $[0, +\infty)$, has a finite mean and is such that,*

$$n\, S(t)\, e^{-\gamma\, t} \ \longrightarrow\ +\infty \quad \text{as } n \to +\infty, \tag{2}$$

for all $t \leq \mathbb{E}[T_{max}]$, with $T_{max} = \max_{i=1,\dots,n} X_0^{(i)}$.

(B) *At each fixed time t, the unpenalized impurity gains $G(t)$ for potential t-splits are i.i.d. samples from a distribution $\mathcal{S}$ with nonzero probability of approaching a finite supremum $G_{\max}(t)$. That is, with large enough $n(t)$, the t-subtree can realize an unpenalized gain arbitrarily close to $G_{\max}(t)$, and consequently a penalized gain $G_{\max}(t)\, e^{-\gamma t}$.*

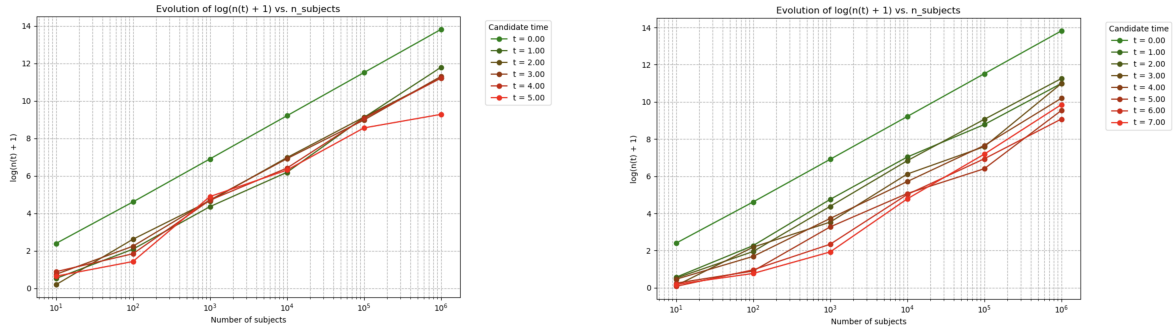
Then, as $n \to \infty$, the Time-penalized Tree estimator is consistent *in the sense that there exists a (nonempty) subset of times t for which all corresponding t-subtrees are consistent.*

4 Validation of Consistency Through Simulations

In this section, we design numerical experiments to *validate* the two key consistency steps outlined previously: demonstrating that the number of subjects effectively reaching a t-node, $n(t)$, grows unboundedly with n when the duration distribution satisfies Assumption (A) and verifying that the depth $K(t)$ of suitably chosen t-subtrees grows unboundedly with $n(t)$ if the distribution of t-gains follows Assumption (B).

We display the evolution of $n(t)$ as the overall sample size n increases, under different choices of the duration distribution (defined at the beginning of Section 4):

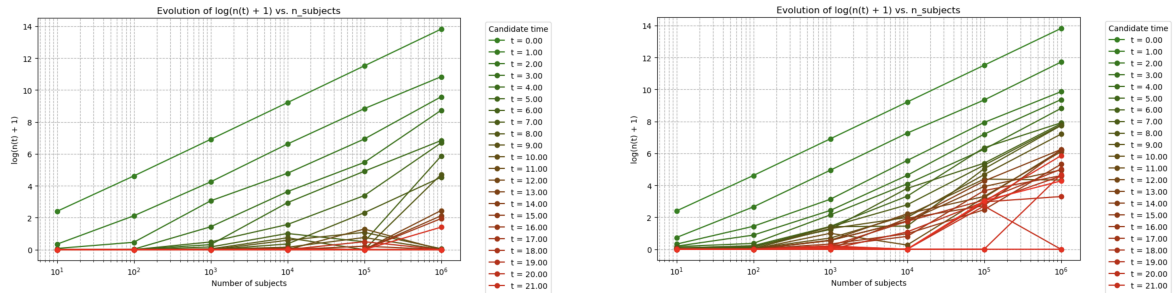
- **Duration distributed over a finite support:** Figure 2 illustrates that when all subjects have the same duration (Dirac distribution ($X_0 = 5$), with a uniform distribution ($\sim$ Uniform$(0, T_{\max} = 7)$), $n(t)$ grows linearly with n for the fixed value $t = t_0$.



(a) Dirac ($x_0 = 5$) **(b)** Uniform(0, 7)

Figure 2: Evolution of $n(t)$ with n for durations distributed over finite supports.

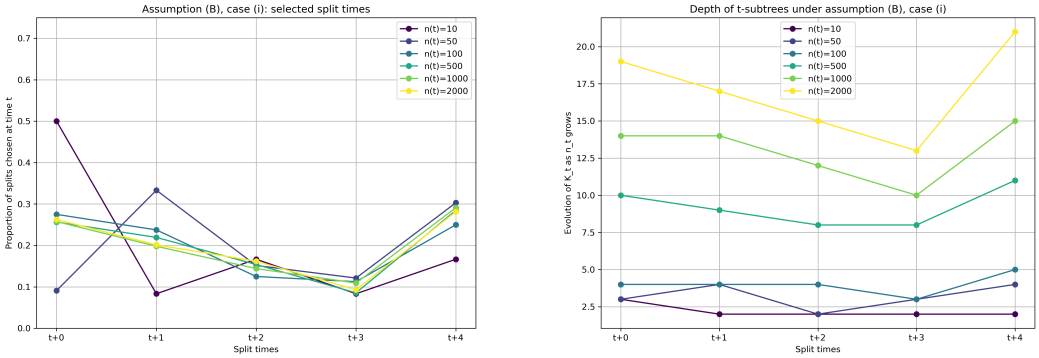
- **Duration distributed over an infinite support:** Figure 3 indicates that, for a log-normal duration, $n(t)$ appears to grow unboundedly with n for small values of t. Regarding if this holds true for any $t < T_{\max}$, is difficult to say for higher t values. In the case of infinite mean, for a Pareto(0, 1), $n(t)$ does not grow unboundedly with n for larger values of t (as $T_{\max} \sim +\infty$).



(a) Lognormal(0, 1) **(b)** Pareto(0, 1)

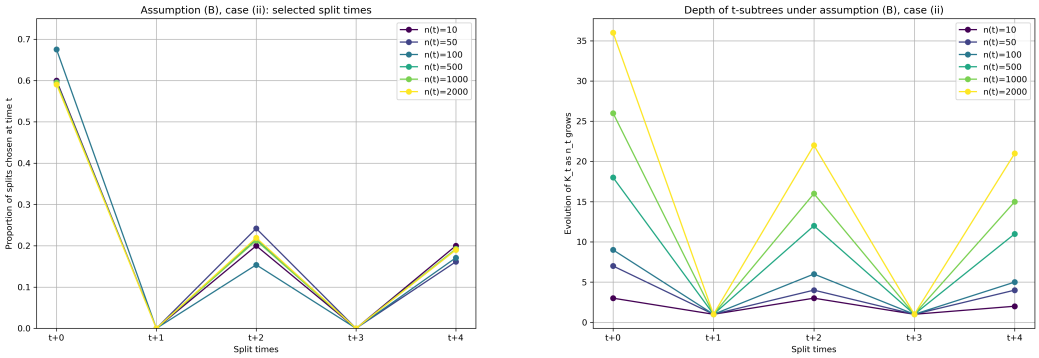
Figure 3: Evolution of $n(t)$ with n for durations distributed over infinite supports.

Next, we analyze the evolution of the depth $K(t)$ of the t-subtrees as the sample size $n(t)$ increases, through three cases reflecting modeling assumptions on unpenalized gains $G(t)$. We show the results of cases (i) i.i.d. Gains across all times, (ii) I.i.d. Gains Within Each Fixed Time t, and (iii) Limit case within a fixed t.



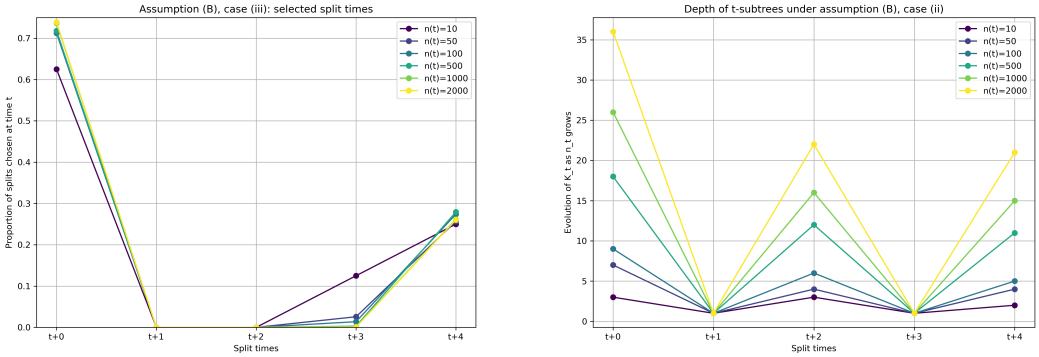
(a) Proportion of split times

(b) Depth K(t)



(c) Proportion of split times

(d) Depth K(t)



(e) Proportion of split times

(f) Depth K(t)

Figure 4: Assumption (B), cases (i), (ii) and (iii): selected splits and t-subtrees depth

Finally, Figure 5 presents results from a full TpT grown on simulated data. The difference with the previous plots is that we see here the full TpT procedure being really applied on a simulated longitudinal dataset, whereas previous graphs were simulating draws from controlled distributions. The overall trend indicates that, globally, the subtree depth $K(t)$ increases with n, thus providing empirical support for the consistency of the tree structure.

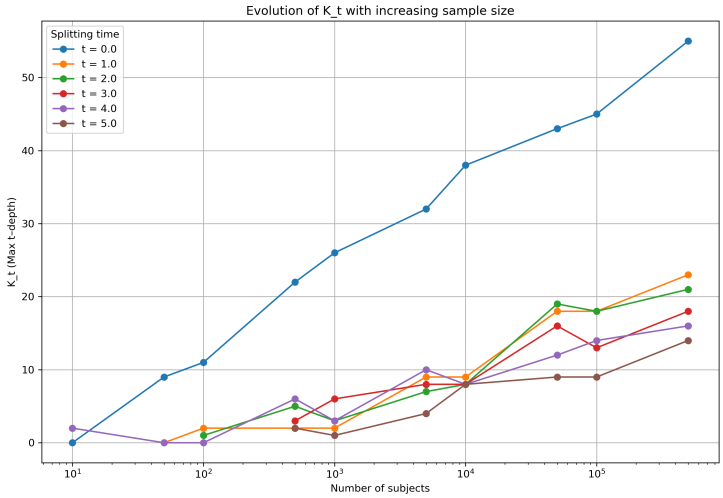


Figure 5: Evolution $K(t)$ as n increases, based on a full TpT grown on simulated data.

5 Conclusion

We established a theoretical framework for the consistency of Time-penalized Trees (TpT) for longitudinal data. Under mild assumptions on subject durations and time-specific gains, every t-subtree in a TpT can eventually amass an unbounded number of observations and grow arbitrarily deep, much like unpruned single-time decision trees. Our proof leverages consistency results for CART-like estimators applied to fixed-time subtrees, showing that the integration of time splits preserves their essential asymptotic properties.

Empirical simulations support these theoretical findings. For duration distributions with finite support or finite mean (e.g., Uniform, Dirac, Log-normal, and Exponential), the effective sample size $n(t)$ increases without bound as the total sample size grows, and the corresponding subtree depths deepen. Even in the borderline case of the exponential distribution, where the expected maximum grows slowly, $n(t)$ still diverges. In contrast, heavy-tailed distributions with infinite mean (such as Pareto with shape $k \leq 1$) do not yield unbounded $n(t)$ for larger t, which underscores the significance of the tail behavior.

Several limitations warrant further investigation. The simulation settings for unpenalized gains were simplified, and interpretation becomes challenging at higher t values due to the steep decay of the exponential penalty. Moreover, while our results illustrate promising asymptotic trends, they primarily address behavior as $n \to \infty$ rather than providing detailed insights into finite-sample performance. Future work should involve comprehensive comparisons with established longitudinal models and consider alternative penalty designs to better capture the complexity of real-world data.

Overall, our study validates TpT as a flexible and interpretable method for analyzing longitudinal data, setting the stage for further research and practical applications.

References

Biau, G. (2012, April). Analysis of a random forests model. *J. Mach. Learn. Res. 13*(null), 1063–1095.

Biau, G., L. Devroye, and G. Lugosi (2008, 09). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research 9*, 2015–2033.

Biau, G. and E. Scornet (2016, Jun). A random forest guided tour. *TEST 25*(2), 197–227.

Breiman, L., J. Friedman, C. Stone, and R. Olshen (1984). *Classification and Regression Trees*. Taylor & Francis.

Fu, W. and J. Simonoff (2016, 12). Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics 18*(2), 352–369.

Grinsztajn, L., E. Oyallon, and G. Varoquaux (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Klusowski, J. M. and P. M. Tian (2024). Large scale prediction with decision trees. *Journal of the American Statistical Association 119*(545), 525–537.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann series in machine learning. Elsevier Science.

Scornet, E. (2016). On the asymptotics of random forests. *Journal of Multivariate Analysis 146*, 72–83. Special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces.

Scornet, E., G. Biau, and J.-P. Vert (2015). Consistency of random forests. *The Annals of Statistics 43*(4), 1716 – 1741.

Valla, M. (2024). Time-penalised trees (tpt): introducing a new tree-based data mining algorithm for time-varying covariates. *Annals of Mathematics and Artificial Intelligence*.

Yao, W., H. Frydman, D. Larocque, and J. Simonoff (2020). Ensemble methods for survival function estimation with time-varying covariates.

Données manquantes

Clustering et réduction de la dimension avec données manquantes

Yasmine Agliz [1] & Vincent Audigier [1] & Mohamed Nadif [2] & Ndèye Niang [1]

[1] *Laboratoire CEDRIC, CNAM, France,*
[2] *Laboratoire Borelli, Université de Paris Cité, France*
{*yasmine.agliz ; vincent.audigier ; n-deye.niang_keita*}*@cnam.fr ;*
{*mohamed.nadif*}*@u-paris.fr*

Résumé.

Dans le cadre de la classification en grande dimension, nous nous intéressons à la méthode Reduced K-means (RKM) de subspace clustering. Cependant, RKM n'est pas adaptée dans le cas de données manquantes. Nous proposons deux méthodes inspirées d'approches couramment utilisées pour la classification avec données manquantes : les approches directes à travers la méthode K-POD et l'imputation multiple (MI). RKM fournit les partitions ainsi que leur sous espace de représentation associés rendant donc nécessaire une phase supplémentaire d'agrégation des sous espaces associées, que nous proposons d'effectuer avec l'AFM. Les deux méthodes sont évaluées sur des données simulées sous les mécanismes MCAR (Missing Completely At Random) et MAR (Missing At Random), avec différents taux de données manquantes. L'évaluation se fait à travers divers scénarios en fonction de l'équilibre des classes, de leur tailles et leur séparation. Les résultats montrent que les deux méthodes parviennent à retrouver les classes initiales et les sous-espaces correspondants dans les divers scénarios. Cependant, la méthode MI-RKM nécessite plus de temps pour l'imputation des données et l'agrégation des paramètres, ce qui accroît la complexité computationnelle.

Mots-clés. Données manquantes, Subspace clustering, Imputation multiple, Analyse factorielle multiple.

Abstract. In the context of high-dimensional clustering, we are interested in the Reduced K-means (RKM) subspace clustering method. However, RKM is not suitable for missing data. We propose two methods inspired by commonly used approaches for clustering with missing data: direct approaches through the K-POD method and multiple imputation (MI). RKM provides the partitions and their associated representation subspaces, thus necessitating an additional phase of aggregation of the associated subspaces, which we propose to perform with AFM. Both methods are evaluated on simulated data under the MCAR (Missing Completely At Random) and MAR (Missing At Random) mechanisms, with different rates of missing data. The evaluation is carried out through various scenarios, depending on the balance of the clusters, their sizes and their separation. The results show that both methods succeed in recovering the original clusters and corresponding subspaces in the various scenarios. However, the MI-RKM method requires more time for data imputation and parameter aggregation, which increases computational complexity.

Keywords. Missing data, Subspace clustering, Multiple imputation, Multiple Factor Analysis.

1 Introduction

En classification non-supervisée en grande dimension, il est courant d'avoir des variables redondantes ou non pertinentes, et parfois manquantes. Pour traiter les variables redondantes et non pertinentes, on peut notamment avoir recours aux méthodes de *Subspace clustering*, qui reposent sur une sélection de variables [1], ou sur une combinaison de variables via des composantes. Dans notre étude, nous nous restreignons aux méthodes basées sur les composantes, et plus particulièrement à la méthode du Reduced K-means (RKM) [2]. Cependant, cette méthode n'est pas adaptée aux données manquantes. Parmi les approches classiquement envisagées pour résoudre ce problème, deux familles de méthodes se distinguent : les approches directes [3, 4] et l'imputation multiple (MI). Les approches directes consistent à définir la fonction objectif d'une méthode de classification non-supervisée uniquement sur les valeurs observées. L'imputation multiple consiste elle à générer dans un premier temps plusieurs versions plausibles du jeu de données. Les paramètres de la méthode de classification non-supervisée sont alors estimés à partir de ces différents jeux de données et sont ensuite agrégés [5]. Dans le cadre du RKM, ces paramètres consistent en la partition des observations, ainsi que le sous-espace dans lequel sont représentés les centres de gravité de chacune des classes.

Nous nous plaçons sous l'hypothèse standard de données manquantes au hasard (MAR) [6] et envisageons ces deux approches de gestion des données manquantes pour le RKM. La gestion des données manquantes par méthode directe, inspirée de l'algorithme K-POD [3], appelée RKPOD, a déjà fait l'objet de précédents travaux [7]. En ce qui concerne l'imputation multiple dans nos travaux précédents, si l'agrégation des partitions a été facilement obtenue par consensus de partitions, le cas est moins évident pour l'agrégation des sous-espaces. Dans ce travail, nous complétons l'approche d'imputation multiple par l'agrégation des sous-espaces à travers la méthode d'Analyse Factorielle Multiple (AFM).

La section 2 présente la méthode RKM dans le cadre complet, puis nous détaillons les deux approches de gestion des données manquantes : d'une part, la méthode directe, qui reformule le problème du RKM pour ne dépendre que des valeurs observées (RKPOD), et d'autre part, notre approche reposant sur l'imputation multiple (MI-RKM). La section 3 décrit le plan de simulation utilisé pour comparer ces méthodes ainsi que les résultats obtenus.

2 Reduced K-means avec données manquantes

La méthode Reduced K-means est une adaptation de la méthode des K-means au contexte du *Subspace clustering*. Elle consiste en la minimisation du critère suivant :

$$\mathcal{C}_{RKM} = ||\mathbf{X} - \mathbf{UFA}^\top||_F^2 \tag{1}$$

où $\mathbf{X}$ de taille $(I \times J)$ correspond à la matrice de données, $\mathbf{U}$ $(I \times c)$ est la matrice d'appartenance des observations aux c classes, composée de 0 et 1, $\mathbf{F}$ $(c \times q)$ est la matrice des centroïdes dans l'espace réduit de dimension q, $\mathbf{A}$ $(J \times q)$ est une matrice de loadings déterminant la contribution de chaque variable à la structure en classe des observations et $||.||_F$ la norme de Frobenius. L'optimisation du critère (1) s'effectue en alternant

entre la recherche de la partition $\mathbf{U}$, obtenue par K-means et la mise à jour du sous-espace $\mathbf{A}$, obtenues par décomposition en valeurs singulières. La matrice des centroïdes $\mathbf{F}$ est calculée à partir des matrices $\mathbf{U}$ et $\mathbf{A}$ [8].

2.1 Méthode directe : RKPOD

La méthode RKPOD redéfinit le critère (1) pour s'accommoder des données manquantes, en utilisant l'opérateur de projection suivant :

$$[P_\Omega(\mathbf{X})]_{ij} = \begin{cases} x_{ij} & \text{si } (i,j) \in \Omega \\ 0 & \text{si } (i,j) \in \Omega^c \end{cases}$$

où $\Omega \subset \{1,\ldots,I\} \times \{1,\ldots,J\}$ est le sous-ensemble des indices des valeurs observées. Ce nouveau critère vise à identifier les centroïdes des observations dans le sous-espace, en s'appuyant sur les valeurs observées de $\mathbf{X}$:

$$\mathcal{C}_{RKPOD} = \|P_\Omega(\mathbf{X}) - P_\Omega(\mathbf{UFA})^\top\|_F^2 \tag{2}$$

Son optimisation est réalisée à l'aide d'un algorithme de Maximisation-Minimisation (MM) [9]. Cet algorithme consiste à se donner un premier jeu de paramètres, puis à identifier une fonction majorante de la fonction de coût en ce jeu de paramètres. Il s'agit de l'étape de maximisation. Ensuite, cette fonction majorante est minimisée, fournissant un nouveau jeu de paramètres. Il s'agit de l'étape de minimisation. Ces étapes sont répétées jusqu'à convergence. Ici, le critère (2) est majoré par le critère (1), dont la minimisation s'effectue selon l'algorithme de RKM. Cet algorithme converge de façon monotone vers un minimum local dépendant de son initialisation. Cette initialisation consiste ici à définir un premier jeu de paramètres $\mathbf{U}, \mathbf{F}, \mathbf{A}$. Pour cela, on commence par remplacer les valeurs manquantes de $\mathbf{X}$ par une méthode telle que les forêts aléatoires (RF) ou k-plus proches voisins (KNN), puis un algorithme de RKM y est appliqué.

2.2 Imputation multiple : MI-RKM

La méthode MI-RKM suit trois principales étapes : d'abord, le jeu de données est imputé M fois, avec une méthode appropriée, comme RF. Ensuite, l'algorithme RKM est appliqué à chaque jeu imputé fournissant une estimation des paramètres notée $\mathbf{U}_m$, $\mathbf{F}_m$ et $\mathbf{A}_m$ $(1 \leq m \leq M)$. Enfin, les paramètres estimés sont agrégés. Pour l'agrégation des partitions $(\mathbf{U}_m)_{1\leq m\leq M}$, un algorithme de factorisation matricielle non négative est utilisé [5]. L'agrégation des sous-espaces et des centroïdes $(\mathbf{A}_m)_{1\leq m\leq M}$ et $(\mathbf{F}_m)_{1\leq m\leq M}$, quant à elle, repose sur une méthode multibloc : l'AFM [10]. Plus précisément, l'agrégation des paramètres est réalisée selon la procédure suivante :

Agrégation de $(\mathbf{U}_m)_{1\leq m\leq M}$ Pour combiner les M partitions $\mathbf{U}_m$, celles-ci sont représentées sous forme de matrices de connectivité $\mathbf{W}_m$, garantissant une matrice positive et évitant le

problème de *label switching*. Le problème de d'agrégation consiste alors à rechercher la matrice de connectivité $\mathbf{W}$, la plus proche de la matrice de connectivité moyenne $\tilde{\mathbf{W}}$, selon le crièrte suivant :

$$\min_{\mathbf{W}} \|\tilde{\mathbf{W}} - \mathbf{W}\|^2, \quad \tilde{\mathbf{W}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{W}_m$$

Cela se fait en factorisant la matrice $\mathbf{W}$ en un produit matriciel de deux matrices $\mathbf{H}\mathbf{H}^\top$, où $\mathbf{H}$ $(I \times c)$ est une matrice d'appartenance permettant d'obtenir directement la partition agrégée $\overline{\mathbf{U}}$.

Agrégation de $(\mathbf{A}_m)_{1\leq m\leq M}$ L'AFM est appliquée pour agréger les M blocs $\{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_M\}$. L'AFM consiste à appliquer l'ACP sur la matrice définie comme suit :

$$\left[\frac{1}{\lambda_1^{(1)}}\mathbf{A}_1 \quad \frac{1}{\lambda_2^{(1)}}\mathbf{A}_2 \quad \cdots \quad \frac{1}{\lambda_M^{(1)}}\mathbf{A}_M\right]$$

où le poids $\lambda_m^{(1)}$ est la première valeur propre de l'ACP du bloc $\mathbf{A}_m$. Cela permet d'avoir un espace compromis de dimension q, qui définit la matrice de loadings agrégée $\overline{\mathbf{A}}$.

Agrégation de $(\mathbf{F}_m)_{1\leq m\leq M}$ Une première approche pour l'agrégation des matrices de centroïdes $\mathbf{F}_m$ consisterait à les combiner directement par AFM. Toutefois, en raison du problème de *label switching*, cette approche n'est pas envisageable. Par ailleurs, les données étant incomplètes, il n'est pas possible de calculer les coordonnées des centres dans l'espace agrégé $\overline{\mathbf{A}}$. Ainsi, nous proposons plutôt d'appliquer l'AFM aux M blocs correspondants au produit des matrices $\mathbf{X}_m\mathbf{A}_m$ obtenu par chacune des imputations. Cela permet de projeter les individus dans un espace commun $\overline{\mathbf{XA}}$. La matrice des centroïdes $\overline{\mathbf{F}}$ est ensuite calculée grâce aux matrices agrégées $\overline{\mathbf{U}}$ et $\overline{\mathbf{XA}}$, selon :

$$\overline{\mathbf{F}} = (\overline{\mathbf{U}}^\top\overline{\mathbf{U}})^{-1}\overline{\mathbf{U}}^\top\overline{\mathbf{XA}}.$$

3 Simulation

Afin d'évaluer la méthode proposée, nous nous sommes inspirés du plan de simulation de [8]. La matrice des données $\mathbf{X}$ est construite comme suit :

$$\mathbf{X} = \mathbf{U}\mathbf{F}\mathbf{A}^\top + \mathbf{E}_{\text{sim}}.$$

Pour chaque scénario, 100 jeux de données sont générés en se basant sur des matrices fixes $\mathbf{U}$, $\mathbf{F}$ et $\mathbf{A}$, tout en faisant varier $\mathbf{E}_{\text{sim}}$, la matrice de bruit. Tout d'abord, un scénario de référence est généré avec un nombre d'individus $I = 1000$, un nombre de variables $J = 50$ et un espace de représentation des centroïdes de dimension $q = 5$. Les classes y sont bien séparées et équilibrées en terme d'effectifs. Ensuite, trois autres scénarios ont été générés : le premier avec des classes déséquilibrées, le deuxième avec des classes qui se chevauchent,

et un dernier avec un nombre plus élevé de variables de bruit corrélées, ce qui se traduit par des coefficients de loadings correspondants nuls. Pour chaque jeu de données, des données manquantes sont générées avec différents taux (15%, 25% et 40%) selon les deux mécanismes MAR et MCAR. L'évaluation est réalisée en utilisant l'Indice de Rand Ajusté (ARI) pour évaluer si la méthode retrouve avec précision la structure des classes simulée (**U**), et le coefficient RV pour déterminer si la méthode parvient à restaurer les sous-espaces **A** et les centroïdes correspondants **F**.

Scénario	Méthode	*U*	*F*	*A*
Classes équilibrées et séparées	MI-RKM	1.000 (0.000)	0.873 (0.009)	0.931 (0.002)
	RKPOD KNN	0.875 (0.016)	0.839 (0.011)	0.920 (0.006)
	RKPOD RF	0.996 (0.002)	0.894 (0.002)	0.929 (0.004)
Classes déséquilibrées	MI-RKM	1.000 (0.000)	0.757 (0.008)	0.933 (0.003)
	RKPOD KNN	0.601 (0.160)	0.799 (0.050)	0.761 (0.033)
	RKPOD RF	0.997 (0.002)	0.919 (0.001)	0.928 (0.004)
Chevauchement des classes	MI-RKM	0.996 (0.005)	0.890 (0.005)	0.937 (0.005)
	RKPOD KNN	0.491 (0.064)	0.663 (0.035)	0.758 (0.004)
	RKPOD RF	0.989 (0.007)	0.895 (0.003)	0.934 (0.006)
Variables de bruit corrélées	MI-RKM	0.780 (0.037)	0.842 (0.089)	0.851 (0.020)
	RKPOD KNN	0.570 (0.027)	0.586 (0.044)	0.799 (0.004)
	RKPOD RF	0.715 (0.035)	0.771 (0.041)	0.808 (0.004)

Table 1: Comparaison des performances des méthodes RKPOD et MI-RKM sur 100 jeux de données avec différents scénarios, avec l'imputation KNN et RF. Avec 40% de données manquantes, et le mécanisme MCAR. Les valeurs indiquées correspondent à la médiane et à l'écart interquartile (entre parenthèses) de l'ARI pour les matrices **U**, et au coefficient RV pour les matrices **A** et **F**. Les résultats soulignés, avec un ARI inférieur à 0.9, un RV de **A** inférieur à 0.9, ou une grande variabilité interquartile, indiquent des performances moins satisfaisantes.

Dans le cas MCAR avec 40% de données manquantes (Table 1), les résultats montrent qu'avec une imputation par RF, MI-RKM et RKPOD ont des performances similaires pour tous les scénarios, sauf celui avec des variables de bruit corrélées. Lorsque RKPOD utilise l'initialisation par KNN, ses performances se dégradent, avec des médianes d'ARI et de RV plus faibles et avec plus de variabilité, surtout dans les scénarios déséquilibrés. L'initialisation des valeurs manquantes a donc un impact majeur sur les performances du RKPOD. En ce qui concerne le scénario avec des variables de bruit corrélées, les deux méthodes rencontrent légèrement plus de difficultés à retrouver le sous-espace et la partition correspondante. Pour des taux de données manquantes de 15 % et 25 %, nous observons les mêmes résultats, avec un écart moins marqué entre MI-RKM et RKPOD. Pour le mécanisme MAR, les conclusions restent les mêmes.

4 Conclusion

Nous avons proposé, dans ce travail, deux approches de gestion des données manquantes : RKPOD et MI-RKM. La méthode RKPOD permet d'obtenir un sous espace directement lié à la partition, tandis que MI-RKM établit ce lien après imputation grâce à la matrice des

centroïdes. L'approche par imputation multiple se distingue par son adaptabilité, car elle peut être appliquée à d'autres méthodes de *Subspace clustering*, contrairement aux méthodes directes comme RKPOD, spécifique à RKM sur données incomplètes. D'autres méthodes multiblocs, comme STATIS, pourraient être envisagées lorsque le taux de données manquantes est faible et que les classes sont faciles à identifier. Pour le RKPOD, l'initialisation des données manquantes a un grand impact sur les résultats, entraînant la nécessité de multiples lancements et donc une augmentation du temps de calcul. Cependant, en comparaison, la méthode MI-RKM demande également du temps pour l'imputation des données et l'agrégation des paramètres, ce qui ajoute une complexité supplémentaire.

References

[1] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Acm sigkdd explorations newsletter*, 6(1):90–105, 2004.

[2] Geert De Soete and J Douglas Carroll. K-means clustering in a low-dimensional euclidean space. In *New approaches in classification and data analysis*, pages 212–219. Springer, 1994.

[3] Jocelyn T Chi, Eric C Chi, and Richard G Baraniuk. k-pod: A method for k-means clustering of missing data. *The American Statistician*, 70(1):91–99, 2016.

[4] Matthieu Marbac, Mohammed Sedki, and Tienne Patin. Variable selection for mixed data clustering: application in human population genomics. *Journal of Classification*, 37(1):124–142, 2020.

[5] Vincent Audigier and Ndèye Niang. Clustering with missing data: which equivalent for rubin's rules? *Advances in Data Analysis and Classification*, pages 1–35, 2022.

[6] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[7] Yasmine Agliz, Vincent Audigier, and Ndèye Niang. Subspace clustering sur données incomplètes. In *Les 55èmes Journées de Statistique*, 2024.

[8] Yoshikazu Terada. Strong consistency of reduced k-means clustering. *Scandinavian Journal of Statistics*, 41(4):913–931, 2014.

[9] Kenneth Lange, David R Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20, 2000.

[10] Jérôme Pages. Multiple factor analysis: Main features and application to sensory data. *Revista Colombiana de Estadística*, 27(1):1–26, 2004.

Exploration de la Régression Logistique avec Données Manquantes dans les Covariables

Christophe Muller [1], Erwan Scornet [2] & Julie Josse [3]

[1] *INRIA PreMeDICaL Inria-Idesp (Inserm - University of Montpellier) team, christophe.muller@inria.fr*
[2] *Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France, erwan.scornet@sorbonne-universite.fr*
[3] *INRIA PreMeDICaL Inria-Idesp (Inserm - University of Montpellier) team, julie.josse@inria.fr*

Résumé. Cette étude présente une évaluation empirique approfondie de différentes procédures de traitement des valeurs manquantes en régression logistique, en se concentrant non seulement sur la performance en classification mais aussi sur la calibration des probabilités et l'estimation des paramètres. Des avancées théoriques récentes ont établi des résultats de consistance pour les méthodes d'imputation en apprentissage supervisé (Josse et al., 2024; Le Morvan et al., 2021) et démontré des liens entre l'imputation constante et la régularisation ridge en régression linéaire (Ayme et al., 2023). Alors que Verchand and Montanari (2024) ont montré que la régression logistique, avec régularisation par ridge, imputée par la moyenne peut atteindre l'optimalité de Bayes en grande dimension, Lobo et al. (2024) ont prouvé les limitations fondamentales des méthodes de classification linéaire avec valeurs manquantes sous conditions MCAR. À travers des simulations extensives, nous comparons plusieurs procédures incluant l'analyse *pattern-by-pattern*, la méthode d'*Expectation-Maximization* avec Approximation Stochastique (`SAEM`, Jiang et al. (2020)), les méthodes d'imputation constante et les méthodes d'*iterative chained equations*. Nos expériences examinent ces méthodes sous différentes conditions de données. Les résultats principaux révèlent que les approches `SAEM` et *pattern-by-pattern* obtiennent généralement une performance prédictive supérieure.

Mots-clés. Régression logistique, Données manquantes, Méthodes d'imputation, Calibration des probabilités

Abstract. This study presents a comprehensive empirical evaluation of various procedures for handling missing values in logistic regression, focusing not only on classification performance but also on probability calibration and parameter estimation. Recent theoretical advances have established consistency results for imputation methods in supervised learning (Josse et al., 2024; Le Morvan et al., 2021) and demonstrated connections between constant imputation and ridge regularization in linear regression (Ayme et al., 2023). While Verchand and Montanari (2024) proved that ridge-regularized logistic regression with mean imputation can achieve Bayes optimality in high dimensions, Lobo et al. (2024) showed fundamental limitations of linear classification methods with missing values under MCAR conditions. Through extensive simulations, we compare several procedures including pattern-by-pattern analysis, Stochastic Approximation EM (`SAEM`, Jiang et al. (2020)), constant imputation methods,

and iterative chained equations methods. Our experiments examine these methods across different data conditions. Key findings reveal that `SAEM` and pattern-by-pattern procedures generally achieve superior prediction performance.

Keywords. Logistic regression, Missing data, Imputation methods, Probability calibration

1 Introduction et État de l'Art

Les données manquantes constituent un défi majeur en apprentissage supervisé, où l'objectif est de prédire une variable de sortie à partir d'un ensemble de caractéristiques d'entrée. Dans les jeux de données réels, les valeurs manquantes surviennent fréquemment pour diverses raisons telles que les erreurs de mesure, la corruption des données ou la non-réponse dans les enquêtes.

Bien que la majorité des travaux existants sur les données manquantes se soit concentrée sur l'inférence et l'imputation, le problème de la prédiction avec des covariables manquantes présente des défis spécifiques. Le Morvan et al. (2020) ont initié l'étude théorique de ce problème en analysant le cas de la régression linéaire. Leurs travaux ont révélé que même lorsque le processus générateur de données est linéaire, le prédicteur optimal devient non-linéaire en présence de valeurs manquantes, nécessitant ainsi des modèles plus complexes que la structure sous-jacente des données ne le suggère.

Josse et al. (2024) ont étendu ces résultats au cadre général de l'apprentissage supervisé sous l'hypothèse MAR (Missing At Random). Ils ont démontré deux résultats complémentaires : d'une part, l'imputation multiple des données de test permet de construire une procédure consistante, et d'autre part, l'imputation constante avant l'apprentissage est consistante lorsqu'elle est combinée avec un algorithme universellement consistant. Le Morvan et al. (2021) ont généralisé ces résultats sans hypothèse sur le mécanisme de données manquantes, montrant notamment que presque toute fonction d'imputation de classe $\mathcal{C}_\infty$ permet d'atteindre l'optimalité de Bayes.

Pour la régression linéaire, Ayme et al. (2023) ont établi que l'imputation par zéro induit un effet de régularisation ridge en grande dimension sous l'hypothèse MCAR. En revanche, la régression logistique présente des défis distincts. Lobo et al. (2024) ont démontré que ni les approches *pattern-by-pattern* ni l'imputation constante ne peuvent atteindre l'optimalité de Bayes sous MCAR avec des covariables indépendantes. Néanmoins, Verchand and Montanari (2024) ont montré qu'en grande dimension, la régression logistique ridge combinée à l'imputation par la moyenne peut atteindre le risque du prédicteur de Bayes.

Notre travail propose une évaluation empirique approfondie des procédures de traitement des valeurs manquantes en régression logistique. Au-delà des résultats théoriques existants, nous évaluons la performance pratique selon trois critères : la qualité de la classification, la calibration des probabilités prédites, ainsi que l'estimation des paramètres du modèle logistique. Cette évaluation est particulièrement pertinente pour la régression logistique, où

la précision des estimations de probabilité est souvent aussi importante que les décisions finales de classification.

Nos expériences évaluent ces procédures sous divers processus de génération de données et en variant la taille des échantillons d'apprentissage. Les résultats principaux montrent que l'algorithme `SAEM` surpasse les autres procédures pour des données Gaussiennes et atteint le risque de Bayes. La procédure *pattern-by-pattern* se raproche également du risque de Bayes mais nécessite un grand nombre de données d'entraînement.

Par ailleurs, l'imputation par MICE (Van Buuren and Groothuis-Oudshoorn, 2011) avec une seule imputation est efficace pour récupérer les paramètres de la régression logistique, ce qui lui permet d'être performante pour la prédiction et la calibration des probabilités sur un ensemble de test avec peu de valeurs manquantes. Cependant, elle est moins performante sur un jeu de données avec beaucoup de valeurs manquantes. L'imputation multiple MICE est plus robuste et offre de bonnes performances, mais n'atteint pas le même niveau que `SAEM` ou que la procédure *pattern-by-pattern* pour un grand nombre d'échantillons d'entraînement.

2 Cadre Théorique : Régression Logistique avec Valeurs Manquantes

Soit $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n$ un jeu de données de n observations, où $\mathbf{x}_i \in \mathbb{R}^d$ représente le vecteur de caractéristiques et $y_i \in \{0, 1\}$ l'étiquette binaire. La régression logistique modélise la probabilité d'appartenance à une classe par la fonction logistique :

$$P(Y = 1|\mathbf{x}; \boldsymbol{\beta_0}) = \sigma(\boldsymbol{\beta_0}^\top \mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\beta_0}^\top \mathbf{x}}}, \tag{1}$$

où $\boldsymbol{\beta_0} \in \mathbb{R}^d$ représente les paramètres du modèle.

En présence de valeurs manquantes, la matrice de données complète $\mathbf{X} \in \mathbb{R}^{n \times d}$ n'est pas entièrement observée. Nous introduisons la matrice indicatrice $\mathbf{M} \in \{0, 1\}^{n \times d}$ définie par :

$$M_{ij} = \begin{cases} 1 & \text{si } X_{ij} \text{ n'est pas observé (NA)} \\ 0 & \text{si } X_{ij} \text{ est observé} \end{cases}. \tag{2}$$

En pratique, nous avons ainsi accès à la matrice des caractéristiques incomplètes $\widetilde{\mathbf{X}}$ où $\widetilde{\mathbf{X}}_{ij} =$ NA si $M_{ij} = 1$, et $\widetilde{\mathbf{X}}_{ij} = X_{ij}$ sinon.

3 Procédures

Plusieurs procédures de traitement des valeurs manquantes dans les covariables sont testées.

Tout d'abord, les procédures d'imputation constante suivies d'une régression logistique sont simples à implémenter et populaires. Nous investiguons deux variantes : l'imputation de

chaque covariable par sa moyenne (`Mean.IMP`) et l'imputation par 0.5 (`05.IMP`). Lobo et al. (2024) ont démontré que de telles procédures ne peuvent généralement pas atteindre le risque de Bayes.

Ensuite, la procédure d'imputation unique par MICE (Van Buuren and Groothuis-Oudshoorn, 2011) suivie d'une régression logistique (`MICE.IMP`) propose une imputation par la distribution conditionnelle des valeurs manquantes en fonction des valeurs observées. Une variante avec imputation multiple (`MICE.MI.IMP`) réalise plusieurs échantillons de la distribution conditionnelle, apprend un modèle logistique sur chaque jeu de données complété, et moyenne ensuite les estimations de chaque coefficient des régressions logistiques.

Trois autres procédures sont testées. L'algorithme EM avec approximation stochastique (`SAEM`) par Jiang et al. (2020) modélise conjointement les covariables et la réponse par un modèle gaussien logistique. On s'attend à ce que `SAEM` soit performant dans ce contexte. La procédure *pattern-by-pattern* (`PbP`) estime une régression logistique pour chaque *pattern* de données manquantes, soit 2^d modèles pour d covariables. Cette procédure ne peut atteindre le risque de Bayes selon Lobo et al. (2024). Enfin, le cas complet (`CC`) exclut les observations contenant une ou plusieurs données manquantes. Cette méthode estime β mais ne peut être appliquée aux données de test contenant des valeurs manquantes.

Pour certaines procédures, nous considérons également une variante incorporant le masque de données manquantes M comme covariable supplémentaire, notée avec le suffixe M (par exemple, `MICE.IMP.M`).

4 Méthodes d'Évaluation

Notre évaluation des modèles de régression logistique avec valeurs manquantes repose sur deux types de métriques. Tout d'abord, les métriques de prédiction et de calibration mesurent la différence entre $\{y_i\}_{i=1}^n$, les labels réels, et $\{\hat{p}_i\}_{i=1}^n$, les probabilités prédites.

- **Score de Brier :** Moyenne des erreurs quadratiques entre probabilités prédites et résultats réels
$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{p}_i)^2. \tag{3}$$
- **Proportion de prédictions correctes** (seuil à 0.5)
$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\mathbb{1}\{\hat{p}_i > 0.5\} = y_i\}. \tag{4}$$

Ensuite, les métriques d'estimation des paramètres mesurent la différence entre $\boldsymbol{\beta_0}$, le vecteur de paramètres réel, et $\hat{\boldsymbol{\beta}}$, son estimation.

- **Erreur Angulaire :** Angle entre les vecteurs de paramètres réels et estimés
$$\text{Erreur Angulaire} = \arccos\left(\frac{\boldsymbol{\beta_0}^T\hat{\boldsymbol{\beta}}}{\|\boldsymbol{\beta_0}\|\|\hat{\boldsymbol{\beta}}\|}\right). \tag{5}$$

- **Erreur Quadratique Moyenne (MSE) :** Moyenne des différences quadratiques des paramètres

$$\text{MSE} = \frac{1}{d}\|\boldsymbol{\beta_0} - \hat{\boldsymbol{\beta}}\|_2^2, \tag{6}$$

où d est la dimension du vecteur de paramètres.

5 Expérience

5.1 Méthodologie

Des données synthétiques sont générées suivant le modèle logistique 1, avec des covariables provenant d'une distribution Gaussienne multidimensionnelle, pour chaque combinaison de paramètres suivants :

- Dimension des caractéristiques : $d = 5$
- Mécanisme de données manquantes : MCAR
- Proportion de valeurs manquantes : $p \in \{25\%, 50\%\}$
- Corrélation entre variables : $\rho \in \{0, 0.65, 0.95\}$
- Proportion de $y = 1$: $\alpha \in \{0.5, 0.85\}$
- Tailles d'échantillon d'entraînement : $n \in \{500, 1000, 5000, 15000, 50000\}$
- Échantillon de test : 15000 observations

5.2 Résultats Principaux

Un premier résultat de notre étude montre que les méthodes d'imputation constante avec masque (`Mean.IMP.M` et `05.IMP.M`) donnent des performances identiques, ce qui est en accord avec la théorie de Le Morvan et al. (2020) (Proposition 3.1).

L'algorithme `SAEM` se distingue comme la meilleure méthode dans toutes les métriques et toutes les configurations de données, atteignant le risque de Bayes. La méthode *pattern-by-pattern* (`PbP`) est également performante lorsque le nombre d'échantillons d'entraînement est grand, s'approchant du risque de Bayes. Cependant, nous observons également que, en accordance avec la théorie de Lobo et al. (2024), les probabilités estimées par `PbP` ne convergent pas vers les probabilités Bayes $\mathbb{P}[Y = 1|X_{obs(m)} = x_{obs}, M = m]$.

Les imputations par la moyenne (`Mean.IMP(.M)`) sont excellentes pour les cas sans corrélation entre les covariables. Cependant, ces procédures sont moins performantes lorsqu'il y a une corrélation entre les covariables.

L'imputation par MICE (`MICE.IMP`) est efficace pour retrouver les paramètres β, ce qui lui permet d'être performante pour l'estimation et la calibration des probabilités prédites lorsque

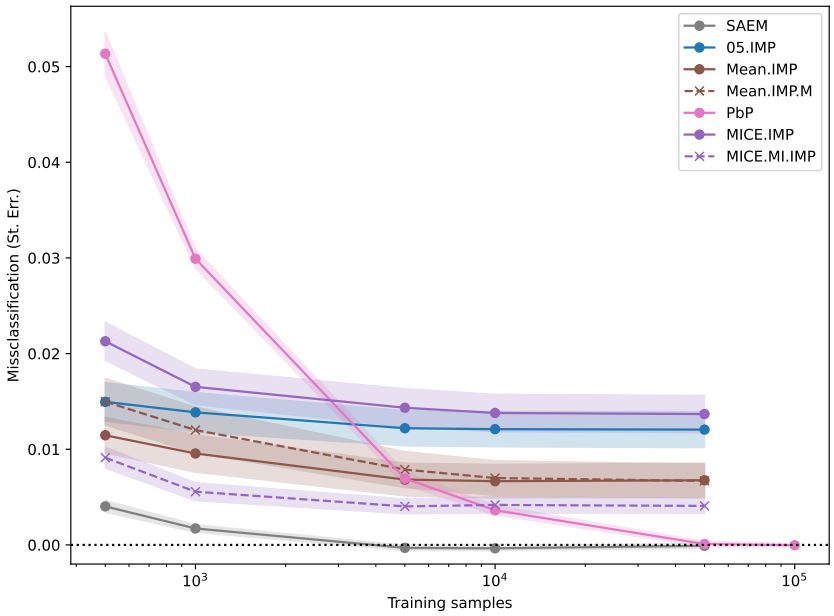


Figure 1: Erreur de classification excédentaire (différence par rapport au classifieur optimal de Bayes) en fonction de la taille de l'échantillon d'entraînement. Les lignes représentent les moyennes sur toutes les configurations expérimentales avec des bandes d'erreur standard.

l'ensemble de test contient peu de valeurs manquantes. Toutefois, elle est moins performante lorsque l'ensemble de test contient beaucoup de valeurs manquantes. L'imputation multiple (`MICE.MI.IMP`) aide à obtenir des résultats plus robustes, mais reste moins performante que `SAEM` ou `PbP` lorsque le nombre d'échantillons d'entraînement est élevé.

L'erreur évaluée en terme de proportion d'observations mal prédites est représentée dans la Figure 1, tandis que l'erreur quadratique moyenne (MSE) pour l'estimation des paramètres β est présentée dans la Figure 2.

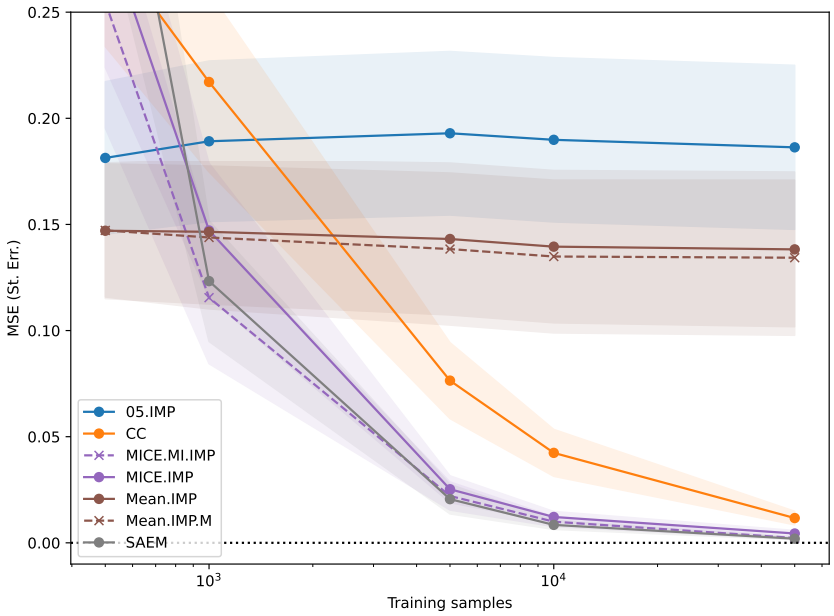


Figure 2: MSE en fonction de la taille de l'échantillon d'entraînement. Les lignes représentent les moyennes sur toutes les configurations expérimentales avec des bandes d'erreur standard.

6 Limitations et Perspectives

Notre analyse s'est limitée à des conditions expérimentales spécifiques. Les extensions à explorer incluent :

- Autres mécanismes de données manquantes : MAR et MNAR, susceptibles de modifier les performances relatives de nos méthodes
- Contextes de grande dimension, où la malédiction de la dimensionnalité pourrait affecter l'imputation et l'estimation des paramètres
- Relations plus complexes entre covariables : dépendances non-linéaires des covariables, structures hiérarchiques

Ces éléments seront développés lors de la présentation des travaux.

Références

Ayme, A., Boyer, C., Dieuleveut, A., and Scornet, E. (2023). Naive imputation implicitly regularizes high-dimensional linear models. In *International Conference on Machine Learning*, pages 1320–1340. PMLR.

Jiang, W., Josse, J., Lavielle, M., Group, T., et al. (2020). Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*, 145:106907.

Josse, J., Chen, J. M., Prost, N., Varoquaux, G., and Scornet, E. (2024). On the consistency of supervised learning with missing values. *Statistical Papers*, 65(9):5447–5479.

Le Morvan, M., Josse, J., Scornet, E., and Varoquaux, G. (2021). What's a good imputation to predict with missing values? *arXiv preprint arXiv:2106.00311*.

Le Morvan, M., Prost, N., Josse, J., Scornet, E., and Varoquaux, G. (2020). Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*, pages 3165–3174. PMLR.

Lobo, A. D. R., Ayme, A., Boyer, C., and Scornet, E. (2024). A primer on linear classification with missing data.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67.

Verchand, K. A. and Montanari, A. (2024). High-dimensional logistic regression with missing data: Imputation, regularization, and universality.

Une introduction à la classification linéaire avec des données manquantes

Angel Reyero Lobo [1], Alexis Ayme [2], Claire Boyer [3] & Erwan Scornet [4]

[1] *Institut de Mathématiques de Toulouse, angel.reyero-lobo@inria.fr*
[2] *École Normale Supérieure*
[3] *Université Paris-Saclay*
[4] *Sorbonne Université*

Résumé. L'apprentissage supervisé avec des données manquantes vise à construire la meilleure prédiction d'une sortie cible à partir d'entrées partiellement observées. Les principales approches pour aborder ce problème peuvent être décomposées en (*i*) des stratégies d'imputation puis de prédiction, qui remplissent d'abord les composants d'entrée vides puis appliquent un prédicteur unique, et (*ii*) des approches Pattern-by-Pattern (P-b-P), où un prédicteur est construit pour chaque motif de données manquantes. Dans cet article, nous analysons théoriquement comment trois classifieurs linéaires classiques, à savoir le perceptron, la régression logistique et l'analyse discriminante linéaire (LDA), se comportent avec des données manquantes complètement aléatoires (MCAR), en fonction de la stratégie (imputation ou P-b-P) utilisée pour gérer les valeurs manquantes. Nous prouvons que les deux approches, l'imputation et le P-b-P, sont mal spécifiées dans un cadre de régression logistique, remettant ainsi en question la pertinence de telles approches pour gérer les données manquantes. Les conditions les plus favorables pour effectuer une classification avec des données manquantes concernent les méthodes LDA P-b-P. Nous fournissons des bornes pour l'excès de risque dans ce cadre, même pour des contextes de grande dimension ou MNAR. Des expériences illustrent nos résultats théoriques.

Mots-clés. Analyse discriminante linéaire (LDA), Régression logistique, Risque de Bayes, Données manquantes complètement aléatoires (MCAR), Données manquantes non aléatoires (MNAR)

Abstract. Supervised learning with missing data aims at building the best prediction of a target output based on partially-observed inputs. Major approaches to address this problem can be decomposed into (*i*) impute-then-predict strategies, which first fill in the empty input components and then apply a unique predictor and (*ii*) Pattern-by-Pattern (P-b-P) approaches,

where a predictor is built on each missing pattern. In this paper, we theoretically analyze how three classical linear classifiers, namely perceptron, logistic regression and linear discriminant analysis (LDA), behave with Missing Completely At Random (MCAR) data, depending on the strategy (imputation or P-b-P) used to handle missing values. We prove that both imputation and P-b-P approaches are ill-specified in a logistic regression framework, thus questioning the relevance of such approaches to handle missing data. The most favorable auspices to perform classification with missing data concern P-b-P LDA methods. We provide finite-sample bounds for the excess risk in this framework, even for high-dimensional or MNAR settings. Experiments illustrate our theoretical findings.

Keywords. Linear discriminant analysis (LDA), Logistic regression, Bayes risk, Missing completely at random (MCAR), Missing not at random (MNAR)

1 Introduction

En raison de la taille croissante des ensembles de données et de l'automatisation de la collecte, les données manquantes sont omniprésentes. Elles peuvent résulter de pannes de capteurs, de non-réponses à des enquêtes ou de l'intégration de sources hétérogènes. Rubin (1976) distingue trois types de données manquantes selon la relation entre les données observées, les données manquantes et le motif de manquabilité: MCAR, MAR et MNAR.

Prédiction La littérature sur les données manquantes se concentre souvent sur l'estimation des paramètres. Néanmoins, la prédiction diffère de l'estimation. Les stratégies imputation-puis-prédiction sont courantes, mais leur analyse reste limitée aux modèles non paramétriques (Le Morvan et al., 2021).Une autre approche consiste à construire un prédicteur par motif de données manquantes (Ayme et al., 2022). Peu d'études théoriques traitent la classification avec données manquantes.

Contributions Nous analysons théoriquement les stratégies d'imputation et de P-b-P en classification linéaire avec données MCAR. Nous montrons que le perceptron et la régression logistique sont mal spécifiés pour ces approches. En revanche, nous prouvons que la LDA en approche P-b-P est optimale, avec des bornes en échantillon fini, même en régime haute dimension ou MNAR.

2 Apprentissage supervisé avec valeurs manquantes

Classification supervisée L'objectif est de prédire une variable cible $Y \in \{-1, 1\}$ à partir d'une observation $X \in \mathbb{R}^d$. Un classifieur $h : \mathbb{R}^d \to \{-1, 1\}$ est évalué via le risque de mauvaise classification :

$$\mathcal{R}_{\text{comp}}(h) = \mathbb{P}(Y \neq h(X)).$$

Le prédicteur de Bayes minimisant ce risque est donné par $h^\star_{\text{comp}}(X) = \text{sign}(\mathbb{E}[Y|X])$. On estime $h^\star_{\text{comp}}$ à partir d'un échantillon $\mathcal{D}_n := \{(X_i, Y_i)\}_{i=1}^n$.

Données manquantes en apprentissage On suppose que X est partiellement observé, avec un masque $M \in \{0, 1\}^d$ où $M_j = 1$ indique une valeur manquante. Notons $X_{\text{obs}(M)}$ et $X_{\text{mis}(M)}$ les sous-vecteurs des valeurs observées et manquantes. La prédiction se fait à partir de $Z = (X_{\text{obs}(M)}, M)$. Le risque associé à un classifieur $h : \mathcal{Z} \to \{-1, 1\}$ est :

$$\mathcal{R}_{\text{mis}}(h) = \mathbb{P}(Y \neq h(Z)),$$

minimisé par le prédicteur de Bayes $h^\star_{\text{mis}}(Z) = \text{sign}(\mathbb{E}[Y|Z])$. En décomposant $h^\star_{\text{mis}}$ par motif de manquabilité $m \in \mathcal{M}$, on obtient :

$$h^\star_{\text{mis}}(Z) = \sum_{m \in \mathcal{M}} h^\star_m(X_{\text{obs}(m)}) \mathbb{1}_{M=m}, \tag{1}$$

où $h^\star_m(X_{\text{obs}(m)}) = \text{sign}(\mathbb{E}\left[Y|X_{\text{obs}(m)}, M = m\right])$. L'objectif est d'estimer chaque $h^\star_m$ à partir de $\mathcal{D}_n = \{(X_{i,\text{obs}(M_i)}, M_i, Y_i)\}_{i=1}^n$.

3 Perceptron

Le perceptron converge sous séparabilité linéaire des données (Novikoff, 1962). Nous étudions ici l'impact des valeurs manquantes sur cette propriété clé.

Lemma 3.1. *La séparabilité linéaire des données complètes n'implique pas celle des données incomplètes.*

Ainsi, l'imputation constante et l'approche P-b-P échouent en général, car la séparabilité pour chaque motif de manquance est nécessaire à la convergence du perceptron.

Lemma 3.2. *Si l'approche P-b-P avec classifieurs linéaires n'est pas Bayes-optimale, alors l'imputation constante avec classifieurs linéaires ne l'est pas non plus.*

4 Régression logistique

Nous analysons ici l'impact des valeurs manquantes sur le modèle de régression logistique.

Assumption 1 (MCAR)**.** $M \perp\!\!\!\perp (X, Y)$.

Assumption 2 (Modèle Logistique)**.** *Il existe* $\beta^\star \in \mathbb{R}^{d+1}$ *tel que* $\mathbb{P}(Y = 1|X) = \sigma(\beta_0^\star + \sum_{j=1}^d \beta_j^\star X_j)$ *avec* $\sigma(t) = 1/(1 + e^{-t})$.

Un modèle P-b-P utilisant une régression logistique spécifique à chaque motif de manquance peut sembler naturel. Cependant, nous montrons que cette approche est mal spécifiée sous MCAR.

Proposition 4.1. *Sous Hypothèses 1 et 2, si les* X_j *sont indépendants à support non borné et* $\mathbb{E}\left[\exp(\beta_j^\star X_j)\right] < \infty$*, alors pour tout* m*, la validité d'un modèle logistique sur* $X_{\text{obs}(m)}$ *impose* $\beta_j^\star = 0$ *pour tout* $j \in \text{mis}(m)$.

Ainsi, le modèle logistique ne peut être valide à la fois sur les données complètes et incomplètes, sauf si les variables absentes sont inutiles dans le modèle initial. Cela montre l'inefficacité des stratégies P-b-P et d'imputation constante. Contrairement à la régression linéaire (Le Morvan et al., 2020; Ayme et al., 2022), la non-linéarité de la régression logistique empêche la préservation de la structure conditionnelle sous MCAR :

$$\mathbb{E}[Y|X_{\text{obs}(M)}, M] \neq \sigma\Big(\mathbb{E}[\beta_0^\star + \sum_{j=1}^d \beta_j^\star X_j|X_{\text{obs}(M)}]\Big).$$

Ainsi, la modélisation de $Y|X$ seule ne suffit pas et nous explorons une autre approche dans la section suivante.

5 Analyse discriminante linéaire

Notre analyse du LDA repose sur l'hypothèse gaussienne suivante :

Assumption 3 (LDA équilibré)**.** *Soit* Σ *une matrice symétrique semi-définie positive de taille* $d \times d$*. On suppose* $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1)$ *et* $X|Y = k \sim \mathcal{N}(\mu_k, \Sigma)$ *pour* $k \in \{-1, 1\}$*.*

Dans le cas complet, le classifieur de Bayes s'écrit :

$$h^{\star}_{\text{comp}}(x) = \text{sign}\Big((\mu_1 - \mu_{-1})^{\top}\Sigma^{-1}(x - \frac{\mu_1 + \mu_{-1}}{2})\Big). \tag{2}$$

Sous MCAR, le classifieur Bayes optimal PbP est donné par :

Proposition 5.1 (Prédicteur de Bayes avec LDA et MCAR)**.** *Sous Hypothèse 1 et 3, on a :*

$$h^{\star}_{m}(x_{\text{obs}(m)}) = \text{sign}\Big((\mu_{1,\text{obs}(m)} - \mu_{-1,\text{obs}(m)})^{\top}\Sigma^{-1}_{\text{obs}(m)} \times (x_{\text{obs}(m)} - \frac{\mu_{1,\text{obs}(m)} + \mu_{-1,\text{obs}(m)}}{2})\Big).$$

Cette optimalité contraste avec le perceptron et la régression logistique, où la stratégie P-b-P était sous-optimale.

Enfin, bien que l'imputation constante soit optimale pour certaines formes de Σ, elle est en général sous-optimale :

Proposition 5.2 (Sous-optimalité de l'imputation constante)**.** *Sous Hypothèses 1 et 3 avec* $d \geq 3$*, l'imputation constante est Bayes optimale pour tout* μ_{-1} *et* μ_1 *si et seulement si* Σ *est diagonale. L'optimum est alors* $\alpha_j = (\mu_{1,j} + \mu_{-1,j})/2$*.*

Nous établissons des bornes de excès de risque pour le classifieur LDA avec des données MCAR (Missing Completely at Random), montrant que ce risque diminue à mesure que la taille de l'échantillon n et la dimension d croissent à des taux prescrits. Nous étendons également notre analyse à des scénarios spécifiques MNAR (Missing Not at Random).

6 Conclusion

Dans cet article, Lobo et al. (2024), nous analysons les deux principales méthodes permettant de gérer les valeurs manquantes en classification : les méthodes Pattern-by-Pattern (P-b-P) et d'imputation. Nous prouvons que les stratégies P-b-P et d'imputation sont théoriquement déficientes pour les

classifieurs perceptron ou régression logistique. En revanche, dans le cadre de l'analyse discriminante linéaire (LDA), la méthode P-b-P est Bayes optimale dans tous les cas, tandis que l'imputation ne l'est que pour les matrices de covariance diagonales.

References

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. Near-optimal rate of consistency for linear models with missing values. In *International Conference on Machine Learning*, pages 1211–1243. PMLR, 2022.

Marine Le Morvan, Nicolas Prost, Julie Josse, Erwan Scornet, and Gaël Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*, pages 3165–3174. PMLR, 2020.

Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. What'sa good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34:11530–11540, 2021.

Angel D Reyero Lobo, Alexis Ayme, Claire Boyer, and Erwan Scornet. A primer on linear classification with missing data, 2024. URL `https://arxiv.org/abs/2405.09196`.

Albert BJ Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pages 615–622. New York, NY, 1962.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 12 1976. ISSN 0006-3444. doi: 10.1093/biomet/63.3.581. URL `https://doi.org/10.1093/biomet/63.3.581`.

Données omiques

Inferring cellular heterogeneity with mixture models for DNA methylation rates

Hugo Barbot [1] & David Causeur [2] & Yuna Blum [3] & Magali Richard [4]

[1] *IRMAR - UMR CNRS 6625, France, hugo.barbot@institut-agro.fr*
[2] *IRMAR - UMR CNRS 6625, France, david.causeur@institut-agro.fr*
[3] *IGDR - UMR CNRS 6290, France, yuna.blum@univ-rennes.fr*
[4] *TIMC - UMR CNRS 5525, France, magali.richard@univ-grenoble-alpes.fr*

Résumé. L'hétérogénéité de la composition en types cellulaires d'échantillons biologiques est un marqueur important de la progression d'une maladie, utile pour son diagnostic. Cette composition cellulaire est cependant difficile à évaluer à partir de profils moléculaires d'un échantillon composite, la contribution de chaque type cellulaire aux signaux observés étant inconnue. La déconvolution cellulaire vise à estimer les proportions des différents types cellulaires à partir de ces profils moléculaires. Deux types de données omiques sont fréquemment employés dans cet objectif: l'expression des gènes et le taux de méthylation de l'ADN. La déconvolution cellulaire s'appuie sur l'hypothèse que le profil moléculaire de l'échantillon composite peut être approché par une somme pondérée de profils moléculaires spécifiques des mêmes gènes pour chaque type cellulaire considéré, les poids étant les proportions inconnues de ces types cellulaires. La plupart des méthodes statistiques utilisées pour la déconvolution cellulaire sont basées sur des extensions de l'algorithme des moindres carrés ordinaires, sous les contraintes de positivité et de somme à un sur les coefficients du mélange. L'utilisation de cet algorithme suppose implicitement l'indépendance, l'homoscédasticité et la normalité des erreurs résiduelles, conditions sous lesquelles il offre des garanties d'optimalité. Dans le cas présent, chacune de ces trois hypothèses est discutable. D'une part, la nature intrinsèque des données omiques requiert des modèles mieux adaptés à leur sur-dispersion. D'autre part, la structure de dépendance induite par le réseau de régulation des gènes est très forte.

Le but de ce travail est de proposer un cadre statistique respectant les caractéristiques inhérentes des données de méthylation de l'ADN. L'analyse exploratoire des distributions des taux de méthylation pour chaque type cellulaire à l'échelle du génome révèle en effet l'existence d'une structure latente en groupes des gènes, chaque composante de cette structure pouvant être associée à un modèle spécifique de déconvolution cellulaire. Nous proposons un mélange de modèles de régression Bêta estimé par maximisation sous contraintes de la vraisemblance grâce à l'utilisation d'un algorithme EM. L'application à des données *benchmark in vitro*, générées à partir de populations cellulaires isolées présentes dans le cancer du pancréas [1], révèle que les modèles de déconvolution cellulaire spécifiques à chaque composante du mélange sont en général très différents. De plus, la sélection des gènes constitutifs de la meilleure composante conduit en général à des améliorations importantes de l'estimation des proportions des types cellulaires. La présentation abordera la question de la sélection des gènes par le choix judicieux d'une composante du mélange pour la déconvolution cellulaire.

Mots-clés. Déconvolution cellulaire, Modèle de mélange, régression Bêta.

Abstract. Cellular heterogeneity in biological tissues reflects progression of disease state

and is therefore useful for improved diagnostic and prognosis. Cellular composition of tissues is however difficult to assess from bulk molecular profiles, with all cells present in the tissue contributing to the recorded signals. Cell deconvolution is a common approach to unravel the heterogeneous molecular profiles observed in bulk tissues, by inferring the underlying relative abundance of individual cell types. Two omics data types are frequently used for cell deconvolution: RNA-seq gene expressions and DNA methylation rates. Cellular deconvolution models assume that bulk omic profiles result from weighted sums of so-called signature cell-specific omic profiles, weights being the unknown proportions of those cell types. Consistently, most statistical methods used for cellular deconvolution are based on extensions of the Ordinary Least Squares (OLS) optimization algorithm, under nonnegativity and sum-to-one constraints on those unknown mixing coefficients. Using OLS implicitly assumes independence, homoscedasticity and normality of the residual errors, conditions under which OLS optimization guarantees optimal estimation. In cellular deconvolution applied to bulk molecular profile, all three assumptions are highly questionable. Indeed, strong violations of those assumptions may be due to the instrinsic nature of omics data, DNA methylation rates being over-dispersed percentages, or to the dependence structure induced by the gene regulatory network, some key genes being more influent on deconvolution accuracy than others.

The goal of this work is to provide a well defined statistical framework that respects the inherent characteristics of biological data for deconvolution, using DNA methylation rates. Whole-genome cell-type specific distributions of DNA methylation rates actually shows a latent group structure, that can explain poor estimation accuracy when fitting deconvolution models on the whole genome. Therefore, we propose a mixture of constrained beta regression models estimated using an EM algorithm. When applied to a *in vitro* benchmark dataset generated from isolated cell populations of Pancreatic cancer(see [1]), results confirms the marked sensitivity of cell deconvolution to the choice of a latent component in the mixture. Interestingly, deconvoluting using the geneset within the best latent component generally leads to a strong improvement with respect to the whole-genome deconvolution. The presentation will compare different gene selection strategies based on the choice of a latent component for cell deconvolution.

Keywords. Cell deconvolution, Mixture models, Beta regression.

1 A brief introduction to single-omic cell deconvolution

Cell deconvolution data produced for the study of Pancreatic ductal adenocarcinoma (PDAC) are used below for an illustrative purpose. This dataset is generated with the aim of serving as a benchmark reference for comparison of cell deconvolution algorithms (see [1]). Therefore, it is obtained under a strict control of the true proportions of each of the 9 cell types in the composition of each bulk. These data consist of DNA methylation rates at more than 800,000 CpG sites over the genome for each of 30 bulks and correspondingly for $p = 9$ cell types. In order to favor the future simultaneous use of DNA methylation and gene expression data in the cell deconvolution task, those methylation rates are aggregated into gene-level

measurements by averaging over all values at CpG sites in the promoter region of each gene. A profile of $m = 18735$ gene methylation rates is available on $n = 30$ independent bulks. The signature matrix contains the methylation rates of those m genes in $p = 9$ cell types.

The basic principles of cell deconvolution are introduced hereafter, based on a single genomic profile of DNA methylation rates. For $1 \leq i \leq n$ and $1 \leq j \leq m$, let y_{ij} denote the methylation rates of gene j for bulk i. Let $y_i = (y_{i1}, \ldots, y_{im})'$ denote the complete measurements profile for bulk i. The aggregated methylation rates $x_j = (x_{j1}, \ldots, x_{jp})',\ j = 1, \ldots, m$ of all genes for p cell types of interest is also available. In standard so-called supervised cell deconvolution models, the methylation profiles in bulks are assumed to result from a linear combination of the signature cell types, the mixing coefficients of this linear decomposition being the unknown proportions $\beta_i = (\beta_{1i}, \ldots, \beta_{pi})'$ of each cell type within each bulk, up to an additive error term. Consistently, most cell deconvolution approaches are variants of the following constrained least-squares minimization issue:

$$(\hat{\beta}_{0i}, \hat{\beta}_i') \quad = \quad \text{argmin}_{(\beta_{0i}, \beta_i')} \sum_{j=1}^{m} (y_{ij} - \beta_{0i} - x_j' \beta_i)^2,$$

where β_{0i} is an intercept and the coefficients β_i are constrained to lie within the $p-$simplex $\mathcal{S}_p = \{\beta = (\beta_1, \ldots, \beta_p),\ 0 \leq \beta_r \leq 1, \sum_{r=1}^{p} \beta_r = 1\}$. The nonnegativity and sum-to-one constraints on β makes the above constrained optimization issue more challenging, which explains the variety of algorithmic solutions available for this task [2].

Unconstrained least-squares optimization is well-studied and guarantees desirable properties, such as unbiasedness and minimum variance, under the standard assumptions of the linear regression model. However, independence, homoscedasticity and normality of the residual error terms are required to ensure optimality of least squares estimation. In the present situation, all three assumptions are highly questionable. Indeed, whereas in the standard approach of omic data analysis, genes are usually considered as features measured on independent statistical units being different biological samples, in cell deconvolution methods, statistical units are genes and features are bulks. Yet, the distribution of DNA methylation rates notoriously shows overdispersion and is driven by a gene regulatory network that induces a graph-structured stochastic dependence pattern across genes.

2 Beta-regression based cell deconvolution

For gene j, given the $p-$profile $x_j = (x_{j1}, \ldots, x_{jp})'$ of signature DNA methylation rates in the p cell types of interest, it is thereafter assumed that the methylation rates y_i, $i = 1, \ldots, n$, in bulk i are independent and distributed according to a Beta distribution with density [3, 4] :

$$\varphi(y_{ij} \mid x_j) \quad = \quad \frac{\Gamma(\phi_i)}{\Gamma(\mu_i(x_j)\phi_i)\Gamma((1-\mu_i(x_j))\phi_i)} y_{ij}^{\mu_i(x_j)\phi_i - 1} (1 - y_{ij})^{(1-\mu_i(x_j))\phi_i - 1}, \qquad (1)$$

where $\phi_i > 0$, $\mu_i(x_j) = \beta_{0i} + \beta_{1i} x_{j1} + \ldots + \beta_{pi} x_{jp} > 0$, β_{0i} is an intercept parameter and $\beta_i = (\beta_{1i}, \ldots, \beta_{pi})' \in \mathcal{S}_p$ is the vector of proportions of each cell type.

A Cyclic Coordinate Descent (CCD) algorithm is implemented to optimize the log-likelihood of model (1). In order to ensure nonnegativity of estimation, at each update of an estimated proportion parameter, if the marginal maximization of the log-likelihood provides a negative update, then the current value of the proportion parameter is set to zero.

Maximum-Likelihood estimations of the proportions of the nine reference cell types in the 30 bulks of the PDAC study are now calculated, using the above CCD algorithm under the assumption of the beta regression model (1). In the present situation where the true proportions of cell types in each bulk are controlled and therefore can be assumed to be known, cell deconvolution performance is evaluated using Mean-Squared-Errors (MSEs) of estimation, for each bulk over the nine cell types.

Figure 1 (left plot) displays boxplots of MSEs over the 30 bulks for `NNLS`, a standard deconvolution algorithm, and the beta regression estimation, referred to as `BR`. It shows that `BR` is slightly outperformed by `NNLS`, both providing poorly accurate estimates of the reference cell type proportions on average over the 30 bulks. Indeed, the right plot in Figure 1 illustrates this lack of accuracy for bulk 18, for which the MSE is close to the median MSE over the 30 bulks (0.006).

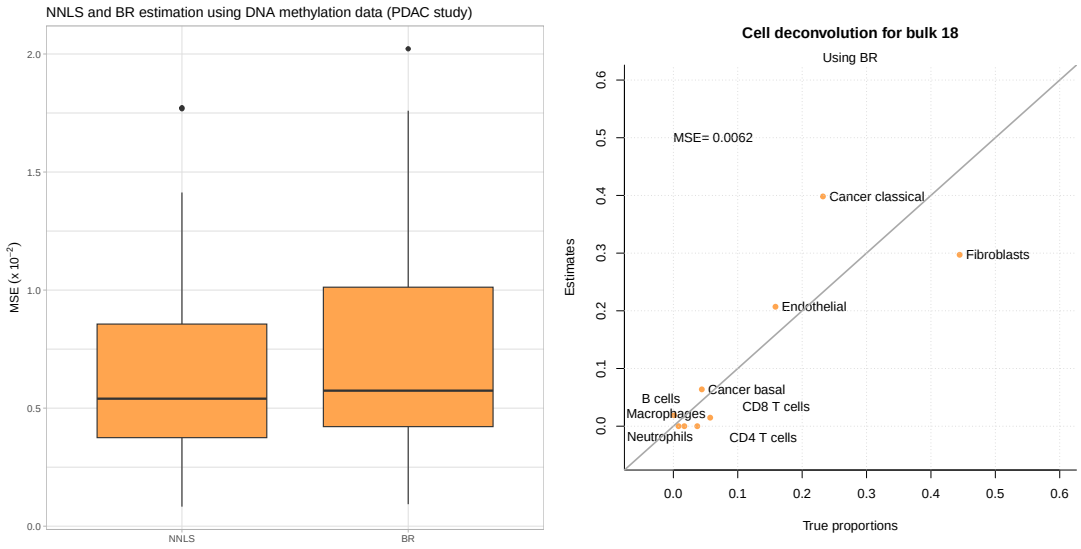


Figure 1: PDAC study. Left plot: estimation accuracy for cell deconvolution using `NNLS` and `BR` - Distribution of MSEs over the 30 bulks. Right plot: comparison of true and estimated proportions, using `BR`, for bulk 18.

Kernel-based estimations of the density functions of the cell type-specific distributions of DNA methylation rates (see left plot in Figure 2) actually questions the assumption of a whole-genome unique deconvolution model. It can be hypothesized that the DNA methylation rates in the signature cell types are not homogeneously distributed according to a single whole-genome Beta distribution, but according to a mixture of distributions whose components are related to latent classes of genes.

Consistently, a mixture model of constrained Beta regression models is assumed thereafter.

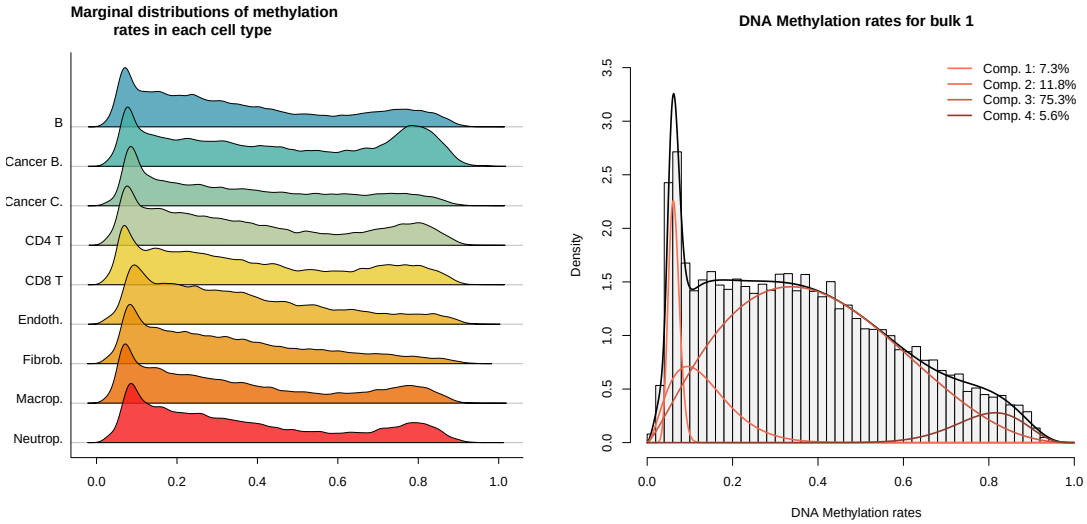


Figure 2: PDAC study. Left plot: marginal distributions of methylation rates in each cell type. Right plot: EM estimation of a 4-component mixture of Beta distributions for DNA methylation rates in Bulk 1, which is the number of component minimizing BIC between 1 and 6.

Let $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ denote a K-latent class variable, distributed according to a multinomial distribution $\mathcal{M}(1, \pi_1, \ldots, \pi_K)$, with $K \leqslant 1,\ \pi_k = \mathbb{P}(C_k = 1) > 0$ and $\sum_{k=1}^{K} \pi_k = 1$. Conditionally on $\mathcal{C}$, the bulk DNA methylation rates Y_{ij}, $j = 1, \ldots, m$ are assumed to be distributed according to a Beta distribution, conditionally on the p-profile of DNA methylation rates $x_{jr},\ j = 1, \ldots, m,\ r = 1, \ldots, p$ in the reference cell types:

$$Y_{ij} | \{\mathcal{C}_{jk} = 1\} \sim \mathcal{B}(\mu_i^{(k)}(x_j) = \beta_{0i}^{(k)} + \beta_{1i}^{(k)} x_{j1} + \ldots + \beta_{pi}^{(k)} x_{jp}, \phi_i^{(k)}).$$

where $\beta_{0i}^{(k)},\ k = 1, \ldots, K$ and $\beta^{(k)} = (\beta_{1i}^{(k)}, \ldots, \beta_{pi}^{(k)}) \in \mathcal{S}_p$ are respectively component-specific intercepts and vectors of proportions for the p cell types in bulk i. The log-likelihood of the above mixture model is maximized using an EM algorithm, in which the M step consists of weighted maximization of componentwise constrained Beta regression models using the CCD algorithm introduced above.

Interestingly, componentwise estimated proportions in early steps of the algorithm offer markedly different viewpoints of the bulk cellular heterogeneity. Both estimated proportions and the corresponding MSEs at the initialisation step are reproduced in Table 1 bor Bulk 1 chosen arbitrarily for illustration. Here, the first two components show a markedly better accuracy than the others compared to the whole genome estimation. More generally, for the 30 bulks, it turns out that at least one of the latent component improves accuracy up to 45% on average.

The presentation will address the dual question of the best stopping criterion for the EM algorithm and the selection of the latent component favoring optimal estimation accuracy.

		Components				
Cell types	Whole genome	1	2	3	4	True proportions
Macrophages	0.00	0.17	0.03	0.00	0.00	0.00
Neutrophils	0.00	0.00	0.00	0.00	0.00	0.00
B cells	0.00	0.00	0.00	0.00	0.00	0.00
CD4 T cells	0.00	0.00	0.00	0.00	0.01	0.01
CD8 T cells	0.00	0.00	0.00	0.00	0.01	0.04
Cancer classical	0.11	0.12	0.12	0.11	0.11	0.05
Cancer basal	0.27	0.12	0.21	0.26	0.14	0.21
Endothelial	0.48	0.25	0.40	0.48	0.53	0.30
Fibroblasts	0.15	0.34	0.23	0.15	0.20	0.38
MSE ($\times 10^{-2}$)	1.08	0.51	0.49	1.07	1.07	

Table 1: PDAC study. Bulk 1. Whole-genome and componentwise cell deconvolution using BR. The true proportions of the 9 reference cell types are provided in the last column. The last row gives the MSE of estimation.

References

[1] Elise Amblard, Vadim Bertrand, Luis Martin Pena, et al. A robust workflow to benchmark deconvolution of multi-omic data. *bioRxiv*, 2024.

[2] Clémentine Decamps, Alexis Arnaud, Florent Petitprez, et al. DECONbench: a benchmarking platform dedicated to deconvolution methods for tumor heterogeneity quantification. *BMC Bioinformatics*, 22(1):473, October 2021.

[3] Timothy J Triche Jr, Peter W Laird, and Kimberly D Siegmund. Beta regression improves the detection of differential dna methylation for epigenetic epidemiology. *BioRxiv*, page 054643, 2016.

[4] Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815, 2004.

Variable selection in transcriptomics data using knockoffs in a classification framework

Julie Cartier [1] & Chloé-Agathe Azencott [2] & Adeline Fermanian [3] & Florian Massip [4]

[1,2,4] *Centre for Computational Biology (CBIO), Mines Paris, Institut Curie, PSL University, INSERM, U1331, France et julie.cartier@minesparis.psl.eu, chloe-agathe.azencott@minesparis.psl.eu, florian.massip@minesparis.psl.eu*
[3] *LOPF, Califrais' Machine Learning Lab, France et adeline.fermanian@califrais.fr*

Résumé. L'émergence de nouvelles technologies de séquençage a facilité l'acquisition de données biologiques en grande quantité qui se sont révélées être des outils indispensables pour mieux comprendre les systèmes biologiques. L'une des manières de tirer parti du potentiel des données de séquençage est de les utiliser pour identifier la relation entre les unités biologiques (ex. les gènes) et les caractéristiques phénotypiques (ex. le diagnostique d'une maladie). Cette question, formulée comme un problème de sélection de variables, reste cependant difficile en raison de la dimension des données ($n < p$) et de leur structure de corrélation. Pour relever ces défis, nous avons étudié l'applicabilité de la procédure knockoff (KO) en nous concentrant sur les données transcriptomiques dans un contexte de classification. Introduite par Candès et al. en 2015, la procédure de sélection de variables KO a montré des résultats prometteurs sur des données biologiques réelles. Cette méthode cherche à identifier les variables prédictives en prenant en compte la structure de corrélation entre les variables tout en contrôlant le taux de fausse découverte. Elle peut également s'appliquer à la grande dimension. Nous avons conduit une étude de simulations à partir des données réelles pour évaluer la pertinence des méthodes récentes dans le cadre de la classification en grande dimension. De plus, nous avons étudié la stabilité de la méthode KO comme mesure de fiabilité de la sélection de variables. Enfin, nous avons analysé les bénéfices d'un schéma d'agrégation de KOs pour réduire l'effet de la stochasticité, intrinsèque à la procédure KO.

Mots-clés. Sélection de variables, Knockoffs, transcriptomique

Abstract. The emergence of new sequencing technologies has facilitated the acquisition of large amounts of biological data, which has proven to be a useful tool for better understanding biological systems. One way to take advantage of the potential of sequencing data is to use them to identify the relationship between biological units (e.g. genes) and phenotypical characteristics (e.g. disease outcomes). This question, formulated as a variable selection problem, remains difficult because of the size of the data ($n < p$) and their correlation structure. To address these challenges, we studied the applicability of the knockoff (KO) procedure focusing on transcriptomic data in a classification setting. Introduced by Candès et al. in 2015, the KO variable selection procedure has shown promising results on real biological data. This method seeks to identify the truly important predictors by overcoming the correlation structure between variables while controlling the false discovery rate even in high dimensional settings. We conducted an extensive simulation study using real data to evaluate the relevance of recent methods in the context of high-dimensional classification. In addition, we studied the stability of the KO method as a measure of the reliability of variable selection. Finally, we analyzed the benefits of a KO aggregation scheme to mitigate the effect of stochasticity, which is intrinsic to the KO procedure.

Keywords. Variable selection, Knockoffs, transcriptomic

1 Résumé des travaux

1.1 Introduction

Grâce aux progrès des technologies de séquençage, il est désormais possible de collecter et d'accéder aux données génomique pour de grandes cohortes de patients, ce qui ouvre la voie à l'utilisation de ces technologies pour de nombreuses applications biotechnologiques et médicales. En particulier, la stratification des patients et la classification des maladies sur la base des résultats d'une expérience de séquençage sont en train de devenir la norme de soins dans plusieurs maladies, et dans le cancer en particulier. Cependant, malgré de nombreuses recherches sur le sujet, la découverte de biomarqueurs à partir de ces données de séquençage reste un défi. Dans le cas de données transcriptomiques, qui mesurent l'expression des gènes, la sélection de biomarqueurs est rendue difficile par leur dimension et la corrélation entre les gènes. En effet, malgré l'augmentation de la taille des cohortes, le nombre de variables explicatives reste beaucoup plus important que le nombre de patients (dans le cas du génome humain, le nombre de variables est de l'ordre de 20,000). De plus, les mécanismes de régulation des gènes induisent une forte corrélation entre ces gènes. La méthode des knockoffs (KOs) est une méthode de sélection de variable introduite par Candes et al. en 2015 [1] puis élargie aux problèmes de grandes dimensions en 2017 [2]. C'est une méthode spécifiquement conçue pour prendre en compte la structure de corrélation des données tout en contrôlant le nombre de faux positifs. De plus, la méthode des KOs possède l'avantage de pouvoir s'adapter à son contexte d'utilisation. Elle peut donc traiter des données en grandes dimensions et des problèmes de classification. Cette méthode a déjà démontré son efficacité avec des données biologiques, en particulier avec des données d'étude d'association pangénomique [3] et avec des problème de régression en grande dimension [2, 4]. Cependant, l'applicabilité des KOs à des données corrélées en grande dimension dans le cadre de la classification reste à établir.
À cette fin, nous avons utilisé des simulations pour, dans un premier temps, étudier comment utiliser les KOs sur nos données et de quelle manière en s'appuyant sur la littérature existante. Nous avons, ensuite, pour évaluer les performances de la méthode des KO et les comparer à d'autres méthodes déjà existantes, à la fois regardé des métriques de performances comme la puissance et le nombre de fausses découvertes faites par les méthodes mais aussi la stabilité comme mesure de fiabilité des résultats.

1.2 Les Knockoffs

Le Modèle-X knockoff La version de la procédure KO que nous utilisons est appelée Model-X knockoffs (MX-KO) et permet de faire de la sélection de variables en grande dimension [2]. Le principe de la procédure KO est de construire des variables fictives 'knockoffs" dont la structure de corrélation est similaire à la structure de corrélation des données originales indépendamment de la variable à prédire. Ce qui se traduit mathématiquement par : Soit $\mathbf{y} \in \mathbf{R}^n$, $\mathbf{y} = (y_1, ..., y_n)^T$ une variable aléatoire représentant la variable à prédire et $X = (X_1, \ ..., X_p)$ l'ensemble des variables $\forall i \in \{1, ..., n\}$, $X_i = (X_{i1}, ..., X_{ip})^T \in \mathbf{R}^p$. Deux hypothèses sont faites. $(X_i, Y_i)_{i\in\{1,...,n\}}$ sont i.i.d et la distribution jointes des variables est connues. La matrice KO, $\widetilde{X}$, de la matrice de variables originales X est construite de telle sorte qu'elle respecte les deux propriétés suivantes :

- Pour tout sous-ensemble $S \subset \{1, ..., p\}$ de variables,

$$(X, \widetilde{X})_{\text{swap}(S)} \stackrel{d}{=} (X, \widetilde{X}), \tag{1}$$

où $(X, \widetilde{X})_{\text{swap}(S)}$ est obtenu en permutant $\widetilde{X}_k$ et X_k dans $(X, \widetilde{X})$ pour tout $k \in S$, et l'égalité des distributions est notées $\stackrel{d}{=}$.

-

$$\widetilde{X} \perp\!\!\!\perp \mathbf{y} | X, \tag{2}$$

avec l'indépendance entre variables notée $\perp\!\!\!\perp$.

Cette nouvelle matrice de données, notée matrice KO, est construite de telle sorte que chaque variable de la matrice originale possède une variable équivalente dans la matrice KO. Ces données synthétiques seront utilisées comme variables de contrôle par le modèle. Ainsi, le modèle sélectionnera les variables dont l'effet sur la prédiction de la variable à prédire sera supérieur à l'effet de la variable KO qui lui est associée. L'objectif étant de sélectionner les variables parce qu'elles sont explicatives et non parce qu'elles sont corrélées à d'autres variables réellement importantes. Dans ce but, une statistique est calculée pour chaque variable. La statistique de la $i^{\text{ème}}$ variable $W_i = w_i((X, \widetilde{X}), \mathbf{y})$, avec w une fonction, doit respecter la propriété suivante :
Pour tout sous-ensemble $S \subset \mathcal{H}_0$ où $\mathcal{H}_0 \subset \{1, ..., p\}$ donne les indices des variables indépendantes de $\mathbf{y}$:

$$w_i((X, \widetilde{X})_{\text{swap}(S)}, \mathbf{y}) = \begin{cases} -w_i((X, \widetilde{X}), \mathbf{y}), \text{ if } i \in S \\ w_i((X, \widetilde{X}), \mathbf{y}) \text{ else.} \end{cases} \tag{3}$$

Finalement, en sélectionnant un sous ensemble de variables $\hat{\mathcal{S}}$ tel que

$$\hat{\mathcal{S}} = \{i : W_i \geqslant \tau\}$$

où $\tau_+ := \min\left\{t > 0 : \frac{1+\#\{i:W_i \leqslant -t\}}{\#\{i:W_i \geqslant t\}} \leqslant q\right\}$ pour un taux de fausses découvertes (FDR) cible q choisi, Candès et al. ont démontré que le FDR de ce sous ensemble de gènes $\hat{\mathcal{S}}$ était contrôlé par le paramètre q [2].

L'agrégation de knockoff En appliquant deux fois la même méthode de construction de variables à la même matrice de données, on peut obtenir des matrices KO différentes en raison de la nature stochastique de ces méthodes. Par conséquent, les variables sélectionnées sont différentes à chaque itération. C'est précisément le problème que la littérature sur l'agrégation des knockoffs cherche à résoudre. Ces méthodes d'agrégation s'appuient sur l'idée que la sélection effectuée après l'agrégation des « connaissances » issues de plusieurs itérations de la procédure KO devrait être plus robuste que la sélection faite à partir d'une seule répétition de la procédure KO. La méthode KOPI (Knockoff-π) est une méthode d'agrégation de KO qui contrôle en probabilité la proportion de faux positifs (FDP) et qui a montré des performances de sélection prometteuses par rapport aux modèles d'agrégation déjà existants [5].

1.3 Simulation des données

Afin d'évaluer les performances de cette méthode, nous avons simulé des données. Plus précisément, nous simulons les variables à prédire $\mathbf{y}$ à partir de vraies matrices de données transcriptomiques X_{data}. Nous simulons $\mathbf{y}$ à partir du modèle linéaire généralisé suivant :

$$\mathbf{y} = \mathbb{1}_{\left(\frac{1}{1+e^{-X_n\beta}} > 0.5\right)}, \tag{4}$$

où $\mathbb{1}$ est la fonction indicatrice. Dans nos simulations : $X_n = \frac{X_{\text{data}} - mean(X_{\text{data}})}{\sigma_{X_{\text{data}}}}$ où $\sigma_{X_{\text{data}}}$ désigne l'écart-type de X_{data}. Pour le vecteur β, nous échantillonnons $k = 10$ variables de manière uniforme dans $\{-1, 1\}$ telles que les $p - k$ autres coefficients du vecteur soient nuls. Nous travaillons principalement avec la matrice réelle X_{CRUKPAP} composée de 369 patients ayant ou non été diagnostiqués d'un cancer du poumon et de l'expression de 749 gènes déterminés comme gènes à risque pour le cancer du poumon par De Biase et al. [6].

1.4 Résultats des expériences

Quelles méthodes pour construire les KOs ? Nous avons comparé différentes méthodes et statistiques utilisées dans la littérature et qui avaient la faculté de s'adapter à nos données transcriptomiques. Pour les méthodes de génération de knockoffs, nous avons choisi de travailler avec l'algorithme LSCIP (Linear Sequential Conditionnal Independant Pairs) [7] qui est une méthode de construction empirique des variables KOs. Cette méthode ne nécessite aucune hypothèse sur la distribution des données. De la même manière, nous avons comparé les performances de sélection obtenues avec différentes statistiques (linéaires et non linéaires) de la littérature et choisi d'utiliser la statistique LCD (Lasso coefficient difference) dans nos expériences [2]. Dans la suite, la procédure KO fera référence à la procédure construite à partir de la méthode LSCIP et de la statistique LCD. Ce sera aussi le cas pour KOPI. Tous les résultats de cette partie ont été validés sur des données simulées à partir de deux nouvelles matrices de données transcriptomiques réelles.

Mesures de performances de la sélection de variables Nous avons comparé les performances de sélection de la procédure KO, à celles obtenues avec d'autres méthodes de sélection de variables déjà utilisées sur ce type de données : une régression logistique avec une pénalisation lasso, et une méthode d'analyse d'expression différentielle (le test de la somme des rangs de Wilcoxon) [8, 9, 10]. Les résultats sont illustrés sur la figure 1 et montrent que le KO permet de retrouver plus de variables non nulles que les autres méthodes pour une proportion de faux positifs donnée. De plus, nous avons comparé les performances de la procédure KO à celles de KOPI et obtenons, à un FDP donné, une puissance moyenne similaire.

Stabilité Afin d'évaluer la stabilité de sélection des différentes méthodes, nous avons répété la sélection de variables sur nos données en perturbant légèrement la matrice de variables. Pour ce faire, à chaque itération, nous ne prenons qu'un sous-ensemble de patients (80%). Puis, nous étudions la fréquence de sélection des différents gènes. Les résultats montrent que la procédure KO n'améliore pas la fréquence de sélection des variables non nulles par rapport au LASSO. La fréquence de sélection dépend du paramètre q (taux de fausse découverte cible) et, pour de petites valeurs de ce paramètre, les KOs tendent à ne sélectionner aucune variable. En revanche, les variables les plus fréquemment sélectionnées sont toujours des variables non nulles du modèle, ce qui n'est pas le cas pour le LASSO. L'étude de la stabilité nous permet aussi d'apprécier la nécessité d'utiliser l'agrégation de KOs pour réduire le nombre de variables alternativement sélectionnées.

1.5 Discussion

En plus des résultats précédents, l'étude de l'applicabilité des KOs à des données transcriptomiques en classification nous a permis d'identifier des limites à son utilisation. La puissance est

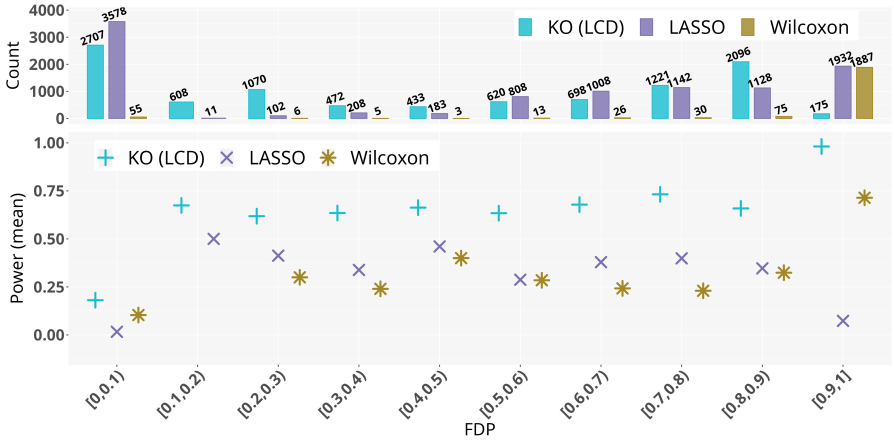


FIGURE 1 – Comparaison des performances de sélection obtenues avec la procédure KO, la régression pénalisée LASSO, et le test de la somme des rangs de Wilcoxon. La moyenne de la puissance est calculée à partir de 100 itérations. Pour chaque otération, $\mathbf{y}$ est simulé linéairement avec $k = 10$ gènes choisis au hasard dans X_{CRUKPAP}. Les histogrammes donnent le nombre de points contenus dans chaque intervallede FDP, i.e. le nombre de points utilisés pour calculer la moyenne.

presque nulle lorsque nous simulons un cas avec de l'interaction entre variables. La puissance diminue lorsque nous augmentons le nombre de variables non nulles. De plus, dans des cas rendus difficiles par de fortes corrélations, la méthode s'avère conservative au point de ne rien sélectionner. Ces limitations soulignent la difficulté de faire de la sélection au niveau de la variable et nous encouragent à considérer l'utilisation des groupes KOs.

Bibliographie

[1] R. F. BARBER et E. CANDÈS. "Controlling the false discovery rate via knockoffs". In : *The Annals of Statistics* 43.5 (2015), p. 2055-2085.

[2] E. CANDÈS et al. "Panning for Gold : 'Model-X' Knockoffs for high dimensional controlled variable selection". In : *Journal of the Royal Statistical Society Series B : Statistical Methodology* 80.3 (2018), p. 551-577.

[3] M. SESIA, C. SABATTI et E. CANDÈS. "Gene hunting with knockoffs for hidden Markov models". In : *Biometrika* 106.1 (2019), p. 1-18.

[4] Asher SPECTOR et Lucas JANSON. "Powerful knockoffs via minimizing reconstructability". In : *The Annals of Statistics* 50.1 (2022), p. 252-276. DOI : 10.1214/21-AOS2104.

[5] A. BLAIN, B. THIRION et O. et al. GRISEL. "False discovery proportion control for aggregated knockoffs". In : *Proceedings of the 37th International Conference on Neural Information Processing Systems* 3417.12 (2024).

[6] M. S. DE BIASE, F. MASSIP et T. T. et al. WEI. "Smoking-associated gene expression alterations in nasal epithelium reveal immune impairment linked to lung cancer risk". In : *Genome Medicine* 16.1 (2024), p. 54.

[7] A. Blain, B. Thirion et O. et al. Grisel. "False discovery proportion control for aggregated knockoffs". In : *NeurIPS*. 2023.

[8] Y. Li, X. Ge et F. et al. Peng. "Exaggerated false positives by popular differential expression methods when analyzing human population samples". In : *Genome Biology* 23.79 (2022).

[9] L. Yang, X. Zhang et J. Chen. "Winsorization greatly reduces false positives by popular differential expression methods when analyzing human population samples". In : *Genome Biology* 25.1 (oct. 2024), p. 282. doi : 10.1186/s13059-024-03230-w.

[10] B P. Hejblum, K. Ba et R. et al. Thiébaut. "Neglecting the impact of normalization in semi-synthetic RNA-seq data simulations generates artificial false positives". In : *Genome Biology* 25.1 (oct. 2024), p. 281. doi : 10.1186/s13059-024-03231-9.

Rank-Based Identification of High-dimensional Surrogate Markers : Application to Vaccinology

Arthur Hughes [1] & Layla Parast [2] & Rodolphe Thiébaut [3] & Boris Hejblum [4]

[1] *INSERM Bordeaux Population Health Research Center, INRIA SISTM, University of Bordeaux and Vaccine Research Institute, INSERM U955, Créteil, France, arthur.hughes@u-bordeaux.fr*

[2] *Department of Statistics and Data Science, University of Texas at Austin, Austin, Texas, United States, parast@austin.utexas.edu*

[3] *INSERM Bordeaux Population Health Research Center, INRIA SISTM, University of Bordeaux, Vaccine Research Institute, INSERM U955, Créteil, and CHU de Bordeaux, Service d'Information Médicale, France, rodolphe.thiebaut@u-bordeaux.fr*

[4] *INSERM Bordeaux Population Health Research Center, INRIA SISTM, University of Bordeaux and Vaccine Research Institute, INSERM U955, Créteil, France, boris.hejblum@u-bordeaux.fr*

Résumé. Dans les essais de vaccins avec un suivi à long terme des participants, il est très important d'identifier des marqueurs de substitution qui permettent de déduire avec précision les réponses immunitaires à long terme. Ces marqueurs offrent des avantages pratiques tels que des preuves précoces et indirectes de l'efficacité des vaccins et peuvent accélérer le développement des vaccins tout en identifiant des biomarqueurs potentiels. Les technologies à haut débit telles que le séquençage de l'ARN sont devenues des outils prometteurs pour comprendre les systèmes biologiques complexes et élaborer de nouvelles stratégies de traitement. Cependant, ces données grandes dimensionnelles présentent des défis statistiques uniques pour les méthodes existantes d'identification de marqueurs de substitution. Nous présentons la méthode Rank-based Identification of high-dimensional SurrogatE Markers (RISE), une nouvelle approche conçue pour les petits échantillons et les contextes à grande dimension typiques des expériences modernes sur les vaccins. RISE utilise un test univarié non paramétrique pour sélectionner les variables des candidats prometteurs, suivi d'une évaluation des marqueurs de substitution sur des données indépendantes. Nos études de simulation démontrent les propriétés souhaitables de RISE, notamment le contrôle du taux d'erreur de type 1 et la puissance empirique dans diverses conditions. En appliquant RISE à un essai clinique de vaccination inactivée contre la grippe, nous avons cherché à identifier les gènes dont l'expression post-vaccinale pourrait servir de substitut à la réponse immunitaire induite. Cette analyse a révélé une signature de gènes dont l'expression combinée un jour après l'injection semble être un substitut raisonnable pour les titres d'anticorps neutralisants à 28 jour après la vaccination. Les voies liées à la signalisation antivirale innée et à la stimulation de l'interféron étaient fortement représentées dans ce substitut dérivé, ce qui permet une interprétation immunologique claire.

Mots-clés. Transcriptomique, Vaccin, Marqueurs de substitution, Grande dimension, Non-paramétrique ...

Abstract. In vaccine trials with long-term participant follow-up, it is of great importance to identify surrogate markers that accurately infer long-term immune responses. These markers offer practical advantages such as providing early, indirect evidence of vaccine efficacy, and can accelerate vaccine development while identifying potential biomarkers. High-throughput technologies like RNA-sequencing have emerged as promising tools for understanding complex biological systems and informing new treatment strategies. However, these data are high-dimensional, presenting unique statistical challenges for existing surrogate marker identification methods. We introduce Rank-based Identification of high-dimensional SurrogatE Markers (RISE), a novel approach designed for small sample, high-dimensional settings typical in modern vaccine experiments. RISE employs a non-parametric univariate test to screen variables for promising candidates, followed by surrogate evaluation on independent data. Our simulation studies demonstrate RISE's desirable properties, including type one error rate control and empirical power under various conditions. Applying RISE to a clinical trial for inactivated influenza vaccination, we sought to identify genes whose post-vaccination expression could serve as a surrogate for the induced immune response. This analysis revealed a signature of genes whose combined expression at 1 day post-injection appears to be a reasonable surrogate for the neutralising antibody titres at 28 days after vaccination. Pathways related to innate antiviral signalling and interferon stimulation were strongly represented in this derived surrogate, providing a clear immunological interpretation.

Keywords. Transcriptomics, Vaccine, Surrogate markers, High-dimensional, Non-parametric
...

1 Introduction

The primary goal of a randomized clinical trial is to evaluate treatment efficacy on a primary outcome, but direct measurement of this outcome can be impractical or unethical. This has led to interest in surrogate markers that can infer treatment effects without direct observation (Christensen et al., 2024). In vaccine trials, surrogate markers are especially valuable, enabling accelerated development during public health emergencies (e.g., COVID-19) and facilitating validation of new vaccines when efficacy trials are infeasible (Gilbert et al., 2024, Plotkin et al., 2010). Transcriptomic data, which capture early gene expression changes post-vaccination, offer promise as surrogates, providing faster endpoints compared to traditional immunological outcomes like antibody titers (Weiner et al., 2015). Studies suggest early gene expression can predict immune responses, but their reliability as surrogates remains uncertain (Querec et al., 2008).

Existing methods for surrogate evaluation, such as those based on Prentice's framework (Prentice, 1989), often rely on parametric assumptions that are difficult to verify in small samples and under complex models. Non-parametric alternatives, though less assumption-dependent, break down with small sample sizes and high-dimensional data (Parast et al., 2015).

To address these challenges, we extend a non-parametric rank-based approach from

Parast et al. (2024) to high-dimensional data, proposing the Rank-based Identification of high-dimensional SurrogatE markers (RISE) method. RISE combines screening of high-dimensional variables with evaluation of their surrogate strength. We demonstrate its utility by applying it to a clinical trial of influenza vaccination to identify potential gene expression surrogates of immunogenicity.

2 Methods

2.1 Notation

Let n represent the total sample size, which is assumed to be small. Denote Y as the primary outcome, $A \in \{1, 0\}$ as a binary treatment indicator (e.g., vaccine or placebo), and $\boldsymbol{S} = (S_1, \ldots, S_p)$ as a set of candidate surrogates, where $p \gg n$. Without loss of generality, assume higher values of Y and S_j are "better," with $j \in \{1, \ldots, p\}$. Using counterfactual notation, each individual has potential outcomes $\left[Y^1, Y^0, \boldsymbol{S}^1, \boldsymbol{S}^0\right]$, where Y^a and $\boldsymbol{S}^a$ denote the primary outcome and surrogate values if the treatment is set to $A = a$.

The observed data consist of n_1 i.i.d. copies of Y^1, $\boldsymbol{S}^1$ for the treatment group and n_0 i.i.d. copies of Y^0, $\boldsymbol{S}^0$ for the control group, with $n = n_0 + n_1$.

2.2 Existing Rank-Based Approach for a Single Surrogate

Parast *et al.* (2024) proposed a rank-based approach to evaluate a single surrogate S_j based on Prentice's definition. A surrogate is valid if the treatment effect on S_j closely matches the treatment effect on the primary outcome Y. These effects are quantified as:

$$U_Y = \mathbb{P}(Y^1 > Y^0) + \frac{1}{2}\mathbb{P}(Y^1 = Y^0), \quad U_{S_j} = \mathbb{P}(S_j^1 > S_j^0) + \frac{1}{2}\mathbb{P}(S_j^1 = S_j^0),$$

with the surrogate's strength measured by $\delta_j = U_Y - U_{S_j}$. Smaller δ_j indicates a stronger surrogate.

The approach uses a non-inferiority test:

$$H_0 : \delta_j \geq \epsilon \quad \text{vs} \quad H_1 : \delta_j < \epsilon,$$

where rejection of H_0 indicates S_j is valid. The Mann-Whitney U-statistics are used to estimate U_Y and U_{S_j}, and δ_j is computed as $\widehat{\delta}_j = \widehat{U}_Y - \widehat{U}_{S_j}$, with standard deviation $\widehat{\sigma}_{\delta_j}$ derived from U-statistic theory.

A one-sided confidence interval for δ_j is given as:

$$\left[-1,\, \widehat{\delta}_j + \Phi^{-1}(1-\alpha)\widehat{\sigma}_{\delta_j}\right],$$

and the p-value is calculated as $p_j = P\left(Z < \widehat{\delta}_j\right)$, where $Z \sim N\left(\epsilon, \widehat{\sigma}_{\delta_j}\right)$. A p-value $p_j < \alpha$ corresponds to the upper confidence bound being less than ϵ.

The choice of ϵ can be adaptive, defined as:

$$\epsilon = \max\left\{0, \widehat{U}_Y - u^*_{\alpha,\beta}\right\}, \tag{1}$$

where $u^*_{\alpha,\beta}$ accounts for the desired power $(1-\beta)$ and significance level α:

$$u^*_{\alpha,\beta} = \frac{1}{2} - \sqrt{\frac{n_0+n_1+1}{12n_0n_1}}\left[\Phi^{-1}(\beta) - \Phi^{-1}(1-\alpha)\right].$$

This ensures that S_j is acceptable as a surrogate if it meets a minimum power threshold.

Overview of RISE

The RISE approach consists of two steps: screening candidate surrogates and evaluating their strength. To prevent overfitting, we use sample splitting to separate the data into independent screening and evaluation sets.

Step 1: Rank-Based Screening

Using the previously described rank-based procedure, we estimate $\widehat{\delta_j}$ and p-values for each candidate surrogate S_j in the screening dataset. To control the false discovery rate (FDR) across multiple tests, we adjust the p-values. The subset $\mathcal{S}$ is selected based on adjusted p-values below a significance level α.

Step 2: Evaluating the Surrogate Set

The selected surrogates in $\mathcal{S}$ are combined into a composite marker:

$$\widehat{\gamma}\mathcal{S} = \sum j \in \mathcal{S} \left|\widehat{\delta}_j\right|^{-1} \bar{S}_j,$$

where $\bar{S}_j$ is standardised, and weights are inversely proportional to $\widehat{\delta_j}$. The rank-based procedure is then applied to $\widehat{\gamma}\mathcal{S}$ in the evaluation dataset. If its p-value is below α, $\widehat{\gamma}\mathcal{S}$ is deemed a valid surrogate for Y.

3 Application to Influenza Vaccination Data

We applied RISE to publicly available data from ImmPort (SDY1276) to identify surrogate markers of immune response to the trivalent inactivated influenza vaccine (TIV) (Bhattacharya et al., 2018). TIV is a seasonal flu vaccine protecting against two influenza A and one influenza B strains. Our analysis focused on young, healthy female volunteers from a 2008–2009 TIV study.

Baseline measurements served as the control, with the primary outcome defined as mean strain-specific neutralizing antibody titres at baseline or day 28 post-vaccination. Gene expression, measured via microarrays, was used to define surrogate marker candidates as baseline and day 1 post-vaccination measurements.

The dataset included observations from 106 unique individuals, with three missing gene expression data at day 1 post-vaccination, resulting in 209 observations: 106 control samples and 103 treated samples. The data were split into screening and evaluation datasets at an $85 : 15$ ratio. A significance level of $\alpha = 0.05$ and a desired power of 0.90 were used for screening, with p-values adjusted using the Bonferroni procedure. The estimated U_Y was 0.95, reflecting a strong neutralizing antibody response, with a non-inferiority threshold of $\epsilon = 0.31$. Among the 10,086 genes, 235 had an adjusted p-value below 0.05. Table 1 presents the top 20 genes sorted by adjusted p-value.

In the evaluation phase, the 235 genes identified in screening were combined into a single predictor, γ_S, using a standardised weighted sum. The estimated U_Y in the evaluation dataset was 0.99, with $\epsilon = 0.15$ for the test powered at 0.90. The value of δ was -0.0063 (95% C.I. $[-0.01, 0.015]$, SD 0.013), with a p-value of $2e - 32$. These results suggest that γ_S is a reasonable surrogate for the neutralising antibody response to 2008-2009 TIV in females. Figure 1 illustrates this, showing a strong Spearman rank correlation ($\rho = 0.79$) between antibody ranks and surrogate ranks.

The biological functions of the 235 genes were examined using DAVID bioinformatics to identify ontological terms which were significantly over-represented in the list. This revealed a significant proportion of these genes to be related to innate antiviral processes, providing a clear immunological interpretation of the signature (Table 2).

In conclusion, we identified a subset of genes whose early post-vaccination expression serves as a promising surrogate for the mid-term immunogenicity of an inactivated influenza vaccine in healthy adult females. This provides a basis for further validation as well as illustrating RISE's practicality as a framework for exploring surrogate markers in clinical studies with high-dimensional candidate markers.

4 Discussion

Surrogate markers offer significant advantages in randomized clinical trials, particularly those assessing vaccine immunogenicity. High-dimensional molecular markers are promising surrogates due to their biological relevance and practical utility. However, existing methods for identifying and validating surrogate markers often fail in high-dimensional contexts, require large sample sizes, or rely on restrictive parametric assumptions. In this study, we introduced RISE—a novel two-step method for identifying and evaluating high-dimensional surrogate markers, applied to a vaccine clinical trial.

Our approach builds on rank-based methodologies, adapting them for high-dimensional settings through univariate testing and dimension reduction, followed by evaluation using independent data. RISE addresses key challenges in evaluating high-dimensional molecu-

lar surrogates, such as the large number of candidate surrogates, limited sample sizes, and the need for false discovery rate control. The initial screening step uses a non-parametric, rank-based univariate test to evaluate whether each variable demonstrates a treatment effect comparable to the primary outcome. This method offers several advantages, particularly in high-dimensional vaccine trials. First, it enables robust inference in small sample scenarios, where assumptions like linearity and normality are hard to verify. Second, being rank-based, it is invariant to data transformations and robust to outliers, which is crucial in high-dimensional data. Finally, by comparing entire rank distributions, this method provides a more comprehensive assessment of surrogate strength. The evaluation step in RISE combines the screened predictors to form a synthetic biomarker.

In applying RISE to a vaccine trial, we aimed to identify early gene expression markers as surrogates for the neutralising antibody response to a trivalent inactivated influenza vaccine (TIV). A signature of 235 genes was identified, with day 1 post-vaccination expression functioning as effective surrogates in the screening data. A weighted combination of these genes was then evaluated on independent data as a potential replacement for the day 28 neutralising antibody response. Biological functions of these genes were explored using DAVID, revealing links to antiviral defence and innate immunity, particularly the interferon response. Further work is needed to assess whether these pathways can predict the vaccine effect across different individuals and vaccines.

While this study advances non-parametric methods for identifying high-dimensional surrogate markers, several limitations must be noted. First, the choice of inferiority margin ϵ for selecting surrogate markers is data-driven but remains ad hoc and difficult to interpret. Sensitivity analyses are recommended to assess how varying ϵ values influence conclusions. The method of combining markers into a single surrogate may also have limited biological interpretation. Additionally, the rank-based approach requires a continuous outcome, which limits its applicability to binary outcomes. Furthermore, RISE identifies statistical surrogates rather than mechanistic ones, and does not account for gene-gene interactions, which are critical in understanding biological pathways. Future improvements should focus on pathway-level analysis and more complex experimental designs, such as longitudinal measurements.

Bibliographie

Christensen R, Ciani O, Manyara AM, Taylor RS. Surrogate endpoints: a key concept in clinical epidemiology.Journal of Clinical Epidemiology.2024;167:111242. doi: 10.1016/j.jclinepi.2023.111242

Gilbert PB, Fong Y, Hejazi NS, et al. Four statistical frameworks for assessing an immune correlate of protection (surrogate endpoint) from a randomized, controlled, vaccine efficacy trial. Vaccine. 2024;42(9):2181–2190. doi: 10.1016/j.vaccine.2024.02.071

Plotkin SA. Correlates of Protection Induced by Vaccination. Clinical and Vaccine Immunology. 2010;17(7):1055–1065. doi: 10.1128/cvi.00131-10

Weiner J, Kaufmann SH, Maertzdorf J. High-throughput data analysis and data integration

for vaccine trials. Vaccine. 2015;33(40):5249–5255. doi: 10.1016/j.vaccine.2015.04.096

Querec TD, Akondy RS, Lee EK, et al. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nature Immunology. 2008;10(1):116–125. doi: 10.1038/ni.1688

Prentice RL. Surrogate endpoints in clinical trials: Definition and operational criteria. Statistics in Medicine. 1989;8(4):431–440. doi: 10.1002/sim.4780080407

Parast L, McDermott MM, Tian L. Robust estimation of the proportion of treatment effect explained by surrogate marker information. Statistics in Medicine. 2015;35(10):1637–1653. doi: 10.1002/sim.6820

Parast L, Cai T, Tian L. A rank-based approach to evaluate a surrogate marker in a small sample setting. Biometrics. 2024;80(1). doi: 10.1093/biomtc/ujad035

Bhattacharya S, Dunn P, Thomas CG, et al. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. Scientific Data. 2018;5(1). doi: 10.1038/sdata.2018.15

Table 1: Screening results from the data application - top 20 genes by adjusted p-values.

Gene	$\boldsymbol{\delta}$ **(95% C.I.)**	$\boldsymbol{\sigma_\delta}$	**Unadjusted p-value**	**B-H Adjusted p-value**
PSME1	-0.037 (-1, -0.011)	0.016	6.3e-109	6.4e-105
NPC2	-0.035 (-1, -0.009)	0.016	1.2e-108	1.2e-104
VAMP5	-0.03 (-1, -0.004)	0.016	7.3e-102	7.4e-98
MYOF	-0.023 (-1, 0.005)	0.017	9.5e-91	9.6e-87
WARS1	-0.026 (-1, 0.002)	0.017	6.4e-88	6.4e-84
UBE2L6	-0.026 (-1, 0.005)	0.018	8.4e-76	8.4e-72
SERPING1	-0.022 (-1, 0.009)	0.019	1.8e-70	1.8e-66
SCO2	-0.017 (-1, 0.015)	0.019	9.7e-68	9.8e-64
STAT1	-0.019 (-1, 0.013)	0.019	1.0e-66	1.0e-62
IFI35	-0.018 (-1, 0.015)	0.020	1.4e-62	1.4e-58
STAT2	-0.016 (-1, 0.018)	0.020	6.3e-60	6.3e-56
GBP1	-0.011 (-1, 0.023)	0.020	2.6e-57	2.6e-53
RHBDF2	-0.008 (-1, 0.027)	0.021	1.6e-52	1.6e-48
GBP2	-0.004 (-1, 0.032)	0.022	2.0e-49	2.0e-45
TYMP	-0.004 (-1, 0.032)	0.022	7.3e-48	7.4e-44
IRF7	-0.003 (-1, 0.033)	0.022	1.7e-47	1.7e-43
XAF1	0.003 (-1, 0.04)	0.022	4.2e-45	4.3e-41
ADAP2	0.009 (-1, 0.045)	0.022	3.7e-44	3.8e-40
ATF5	0.008 (-1, 0.045)	0.023	7.4e-42	7.4e-38
SP140	0.004 (-1, 0.042)	0.023	2.4e-41	2.5e-37

Table 2: Functional annotation of 235 significant genes using DAVID. The table shows functional terms, the number of associated genes, and adjusted p-values for a test of term over-representation compared to a random sample of the same size from the 10,086 aligned genes.

Term	**Number of genes**	**B-H Adjusted p-value**
KW-0051-Antiviral defense	37	3.1e-33
GO:0051607-defense response to virus	43	5.6e-31
KW-0391-Immunity	59	1.1e-27
KW-0399-Innate immunity	45	4.2e-22
GO:0045071-negative regulation of viral genome replication	20	2.1e-19
GO:0045087-innate immune response	43	5.7e-18
GO:0009615-response to virus	24	2.3e-15
hsa05164:Influenza A	23	1.0e-10
GO:0034341-response to type II interferon	11	1.9e-09
hsa04621:NOD-like receptor signaling pathway	22	8.9e-10

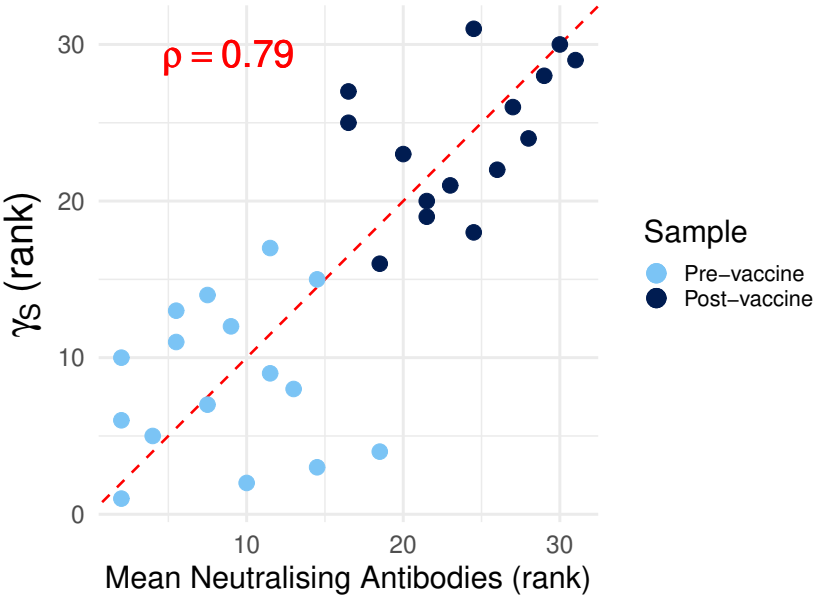


Figure 1: Ranks of the mean neutralising antibodies against the ranks of the constructed 235-gene-combination surrogate marker in the evaluation dataset. The Spearman rank correlation coefficient is 0.79, indicating strong positive correlation.

Inférence post hoc pour l'analyse différentielle de données Hi-C

Élise Jorge[1,2], Pierre Neuvial[3], Nathalie Vialaneix[2] & Sylvain Foissac[1]

[1] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet-Tolosan, France,* {*elise.jorge, sylvain.foissac*}*@inrae.fr*

[2] *Université Fédérale de Toulouse, INRAE, MIAT, 31326 Castanet-Tolosan, France, nathalie.vialaneix@inrae.fr*

[3] *Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS, UPS, F-31062 Toulouse Cedex 9, France, pierre.neuvial@math.univ-toulouse.fr*

Résumé. Les données Hi-C fournissent une information sur l'organisation tridimensionnelle du génome à partir de mesures d'interactions entre positions génomiques le long de la chromatine. Cette structure en trois dimensions a un rôle important dans la régulation de l'expression des gènes. L'objectif de l'analyse différentielle est d'identifier, à partir de réplicats obtenus dans deux conditions biologiques différentes, des régions génomiques qui présentent des différences significatives de structure entre les deux conditions. Ici, nous proposons de nous appuyer sur des outils d'inférence post hoc couplés à de l'analyse différentielle restreinte aux pixels. Il est ainsi possible de quantifier la présence d'interactions différentielles dans des sous-ensembles de pixels arbitrairement choisis permettant ainsi d'identifier les régions génomiques les plus différentielles.

Mots-clés. données Hi-C, génomique 3D, inférence post hoc, tests multiples

Abstract. Hi-C data provide insights into the three-dimensional organization of the genome by measuring interactions between genomic positions along the chromatin. This three-dimensional structure plays a crucial role in regulating gene expression. Differential analysis aims to identify genomic regions that display significant differences in structure between two different biological conditions. Here, we propose to use a post hoc inference strategy on results obtained from pixel-level differential analysis. This makes it possible to quantify signal in arbitrary clusters of pixels and thus to identify differential genomic regions.

Keywords. Hi-C data, 3D genomics, post hoc inference, multiple testing

1 Introduction

Structure de l'ADN et données Hi-C. La chromatine est compactée au sein du chromosome selon une structure hiérarchique, comme illustré sur la figure 1 (gauche). Les données Hi-C sont des données de séquençage haut-débit qui permettent d'obtenir des informations sur l'organisation tridimensionnelle du génome dans la cellule, en mesurant la fréquence d'interactions spatiales entre régions génomiques. L'étude de ces données a permis de montrer

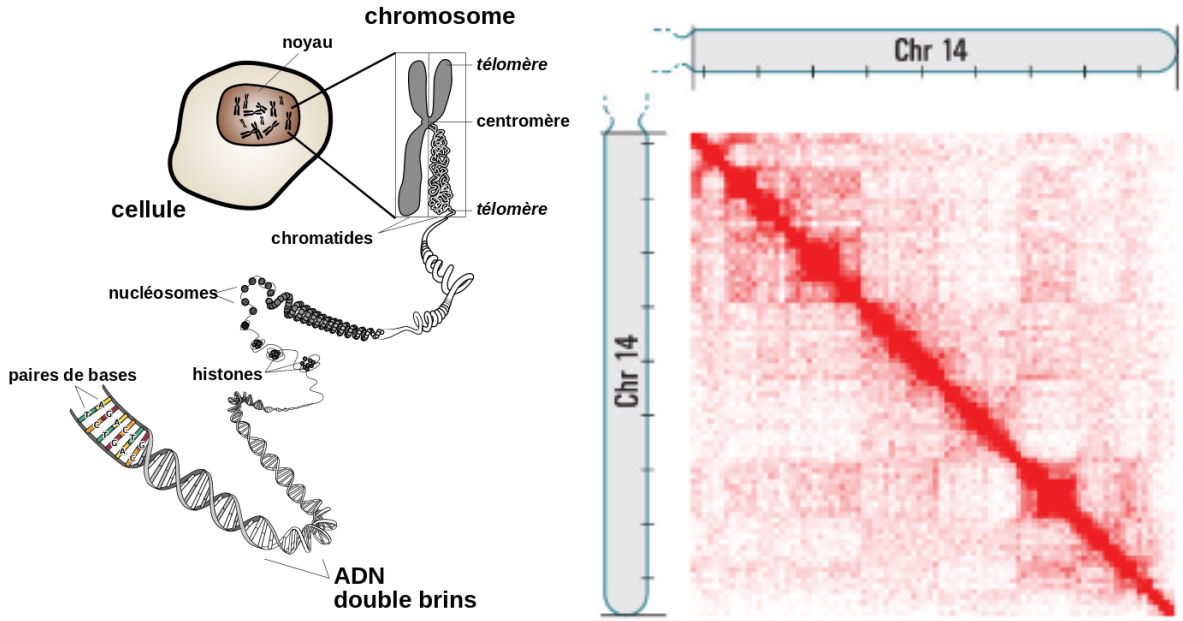


FIGURE 1 – Gauche : Schéma de la compaction de l'ADN en chromosome (« Chromosome fr » par Phrood commonswiki, Wikimedia Commons). Droite : Matrice Hi-C du chromosome 14 de [1].

qu'il existait, le long de la chromatine, des régions génomiques appelées TADs (*Topologically Associating Domains*) au sein desquels les interactions sont fréquentes.

L'apparition de modifications dans cette structure de compaction, par exemple la disparition d'une frontière entre deux TADs impliquant leur fusion, peut avoir un impact majeur sur l'expression des gènes dans la zone considérée. Ces modifications peuvent provoquer des pathologies neurologiques [2] ou des malformations [3].

D'une manière plus formelle, les données Hi-C se présentent sous la forme d'une matrice carrée symétrique dont l'entrée (i, j) – que l'on appellera « pixel » – correspond au nombre d'interactions observées dans l'expérience Hi-C entre les positions génomiques i et j. La figure 1 (droite) représente une telle matrice, dans laquelle l'intensité de couleur est proportionnelle à la valeur du nombre d'interactions du pixel considéré.

Analyse différentielle de données Hi-C. On s'intéresse ici à un problème d'analyse différentielle entre des ensembles de matrices Hi-C obtenues dans deux conditions différentes, $\mathcal{C}_1$ et $\mathcal{C}_2$. L'objectif est d'identifier des régions génomiques qui présentent des différences significatives d'interactions entre ces deux conditions.

Formellement, on possède $r = r_1 + r_2$ matrices de taille $p \times p$ où $M_k^{\mathcal{C}_1}$ $(k = 1, \cdots, r_1)$ (resp. $M_l^{\mathcal{C}_2}$ $(l = 1, \cdots, r_2)$) correspond à la matrice obtenue pour le k-ème (resp. l-ème) réplicat de la condition $\mathcal{C}_1$ (resp. $\mathcal{C}_2$).

L'analyse différentielle peut être réalisée au niveau des pixels, en faisant l'hypothèse $H_0^{(i,j)}$: « *Le nombre moyen d'interactions entre les paires de positions génomiques i et j n'est pas*

différent entre les deux conditions » avec $i, j \in \{1, \cdots, p\}$ et $j \leq i$ (par symétrie, on ne regarde que la partie triangulaire supérieure de la matrice). Ainsi, le test de ces hypothèses permet d'obtenir une unique p-valeur pour chaque pixel (i, j).

Dans un benchmark [4] de méthodes d'analyse différentielle au niveau des pixels, nous avons montré que **diffHic** [5] est une des méthodes fournissant les meilleures garanties statistiques. Cependant, l'interprétation des résultats de ce type de méthode est limitée d'un point de vue biologique car l'information obtenue est ponctuelle, très dispersée sur la matrice et ne peut pas être facilement interprétée en « régions de compaction / décompaction » dans la matrice. De plus, les résultats par pixel ne sont pas facilement et directement généralisables à des ensembles de p-valeurs. Ainsi, nous nous intéressons dans la suite à une manière « d'agréger » l'information fournie par ces résultats afin d'identifier des régions génomiques différentielles.

2 L'inférence post hoc pour l'analyse différentielle

Dans cette partie, nous nous intéressons à la problématique d'extension des résultats de l'analyse différentielle ponctuelle donnant une p-valeur par pixel à une analyse différentielle sur des ensembles de pixels. On montre notamment qu'utiliser l'inférence post hoc peut permettre de quantifier le signal présent dans des sous-ensembles de p-valeurs.

Limites du contrôle du False Discovery Rate. Un grand nombre de tests étant réalisés simultanément – ici, on réalise théoriquement jusqu'à $p(p+1)/2$ tests – on doit ajuster les p-valeurs obtenues pour corriger pour la multiplicité des tests. Les p-valeurs fournies par la méthode **diffHic** sont corrigées à l'aide de la méthode de Benjamini-Hochberg (BH) [6], assurant un contrôle du FDR sur l'ensemble des pixels testés. Si l'on note $\mathcal{H}_0$ l'ensemble des hypothèses nulles et R l'ensemble des hypothèses rejetées par la méthode, le FDR s'écrit comme $\mathbb{E}\left(\mathrm{FDP}(R)\right)$, où $\mathrm{FDP}(R) = \frac{|R \cap \mathcal{H}_0|}{|R| \vee 1}$ est la proportion de faux positifs parmi l'ensemble des hypothèses rejetées par la méthode. Alors, il apparaît que le contrôle global du FDR sur l'ensemble des pixels considérés n'implique par le contrôle de cette quantité sous un sous-ensemble de pixels sélectionnés [7]. Ainsi, il n'est donc pas possible de fournir de garanties sur la présence de faux positifs dans un sous-ensemble de p-valeurs ajustées seulement en utilisant la méthode BH.

Inférence post hoc : objectif et définition. Ici, notre objectif est de donner une mesure de la quantité de signal dans des sous-ensembles de p-valeurs *arbitrairement sélectionnés*. Les méthodes post hoc [7] fournissent précisément une telle garantie. Pour tout ensemble S de p-valeurs, elles fournissent une borne supérieure sur $|S \cap \mathcal{H}_0|$, le nombre de faux positifs dans S. Formellement, on appelle *borne post hoc* [7] une fonction V_α telle que :

$$\mathbb{P}(\forall S, |S \cap \mathcal{H}_0| \leq V_\alpha(S)) \geq 1 - \alpha\,. \tag{1}$$

Si (1) est vérifiée, alors pour tout S la quantité $\gamma_\alpha(S) = 1 - V_\alpha(S)/|S|$ permet de minorer la proportion de vrais positifs $\mathrm{TDP}(S) = 1 - |S \cap \mathcal{H}_0|/S$:

$$\mathbb{P}\left(\forall S, \mathrm{TDP}(S) \geq 1 - \frac{V_\alpha(S)}{S}\right) \geq 1 - \alpha\,. \tag{2}$$

Comme $\gamma_\alpha(S)$ minore la proportion de vrais positifs dans un sous-ensemble d'intérêt S, indépendamment du choix de S, cette mesure peut être utilisée pour comparer des sous-ensembles de p-valeurs arbitrairement sélectionnés.

Utilisation de la borne post hoc de Simes. Si l'on note $p_1, \ldots, p_m$ les p-valeurs correspondant aux m hypothèses nulles, on définit la borne de Simes :

$$V_\alpha^{\mathrm{Simes}}(S) = \min_{1 \leq k \leq m} \left[\sum_{i \in S} \mathbb{1}_{\{p_i > \frac{\alpha k}{m}\}} + k - 1 \right].$$

Sous des hypothèses d'indépendance des p-valeurs ou de dépendance positive (PRDS) [8], on peut montrer [9] que la borne V^{Simes} satisfait l'équation (1) (et est donc une borne post hoc). L'hypothèse PRDS est considérée comme réaliste pour les applications génomiques [10]. En particulier, c'est sous cette hypothèse que le contrôle du FDR par la procédure BH est valable.

Nous proposons donc d'utiliser la borne $V_\alpha^{\mathrm{Simes}}$ dans le cadre de l'analyse différentielle de données Hi-C.

3 Application : changement de conformation durant la différentiation cellulaire

Nous avons implémenté la méthode décrite en section 2, et l'avons testée sur des données Hi-C issues de lignées cellulaires murines [11] pour deux conditions biologiques correspondant à des stades de différenciation cellulaire différents de cellules neuronales (ES : cellules souches embryonnaires et CN : neurones corticaux).

Ici, les ensembles arbitraires ont été définis à partir des données en utilisant une classification ascendante hiérarchique avec contrainte de voisinage [12]. Soient $i, j \in \{1, \cdots, p\}$ avec $j \geq i$. Dans la classification, chaque pixel (i, j) est représenté par $\mathbf{m}_{i,j} := (m_{i,j}^{\mathcal{C}_1,1}, \cdots, m_{i,j}^{\mathcal{C}_1,r_1}, m_{i,j}^{\mathcal{C}_2,1}, \cdots, m_{i,j}^{\mathcal{C}_2,r_1})$ correspondant aux comptages de l'interaction entre les positions génomiques i et j dans tous les réplicats des deux conditions. Nous utilisons l'implémentation **scikit-learn** avec le critère de Ward [13]. Le dendrogramme obtenu est coupé à une hauteur h choisie en cherchant un "coude" dans l'allure de l'évolution du critère de Ward en fonction de l'étape de la classification ascendante hiérarchique. Les bornes post hoc sont calculées sur les classes ainsi obtenues.

Sur la figure 2, nous représentons les résultats obtenus pour le chromosome 18. La partie inférieure de la figure correspond aux p-valeurs ajustées de l'analyse différentielle par pixel

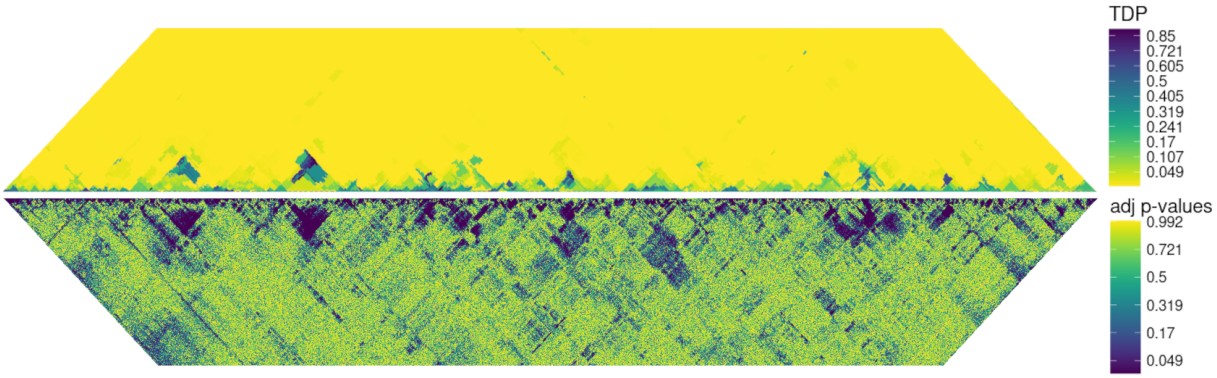


FIGURE 2 – En haut, diagonale de la (demie) matrice des résultats de l'inférence post hoc où chaque classe de pixels est représentée par sa proportion minimale de vrais positifs calculée par l'inférence post hoc. En bas, diagonale de la (demie) matrice des p-valeurs ajustées obtenues par la méthode **diffHic**.

obtenue par la méthode **diffHic**. On distingue des régions dans lesquelles se trouve un nombre important de p-valeurs ajustées contigues significatives mais ces zones sont mélées de p-valeurs ajustées plus fortes. Inversement, des p-valeurs ajustées faibles se retrouvent de manière éparpillée dans des zones isolées de la matrice.

La partie supérieure de la figure représente la proportion minimale de vrais positifs des classes obtenues par classification ascendante hiérarchique. On observe une correspondance entre des régions avec une forte densité de p-valeurs ajustées faibles (partie haute) et des classes possédant une proportion de vrais positifs élevée.

4 Perspectives

Ainsi, l'utilisation de l'inférence post hoc couplée à une approche totalement automatisée de classification ascendante hiérarchique a permis de passer d'une information ponctuelle (les p-valeurs ajustées fournies par **diffHic**) et biologiquement limitée à une information sur des régions génomiques différentielles.

Dans la suite, on s'intéressera à la validation biologique des résultats obtenus, en identifiant par exemple les gènes présents dans les classes dont les proportions de vrais positifs sont les plus élevées.

L'approche que nous avons décrite est en cours d'évaluation sur d'autres jeux de données comme les données [14] correspondant à des cellules murines post-mitotiques. Dans ce jeu de données, on retrouve une condition contrôle (CTCF+) et une condition dans laquelle une déplétion de la protéine CTCF a été provoquée (CTCF-). Dans la condition de traitement (CTCF-), une modification de structure est attendue de par la présence favorisée de la protéine CTCF dans des régions telles que les frontières de TADs. Nous souhaitons utiliser des informations biologiques telles que la présence de pics CTCF afin d'obtenir une validation des classes identifiées par la méthode comme différentielles.

Remerciements

Ce travail est soutenu par le groupe de travail ChrocoNET financé par le métaprogramme INRAE DIGIT-BIO. La thèse d'Élise Jorge est financée par INRAE.

Bibliographie

[1] Erez Lieberman-Aiden, Nynke L. Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M.A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[2] Malte Spielmann, Darío G Lupiáñez, and Stefan Mundlos. Structural variation in the 3d genome. *Nature Reviews Genetics*, 19(7):453–467, 2018.

[3] Darío G Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M Opitz, Renata Laxova, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

[4] Élise Jorge, Sylvain Foissac, Pierre Neuvial, Matthias Zytnicki, and Nathalie Vialaneix. A comprehensive review and benchmark of differential analysis tools for Hi-C data. *Briefings in Bioinformatics*, 2025. Forthcoming.

[5] Aaron T.L. Lun and Gordon K. Smyth. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics*, 16:258, 2015.

[6] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[7] Jelle J Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584 – 597, 2011.

[8] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

[9] Gilles Blanchard, Pierre Neuvial, and Etienne Roquain. Post hoc confidence bounds on false positives using reference families. *The Annals of Statistics*, 48(3):1281–1303, June 2020.

[10] Jelle J Goeman and Aldo Solari. Multiple hypothesis testing in genomics. *Statistics in medicine*, 33(11):1946–1978, 2014.

[11] Boyan Bonev, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L Papadopoulos, Yaniv Lubling, Xiaole Xu, Xiaodan Lv, Jean-Philippe Hugnot, Amos Tanay, and Giacomo Cavalli. Multiscale 3D genome rewiring during mouse neural development. *Cell*, 171(3):557–572.e24, 2017.

[12] Nathanaël Randriamihamison, Nathalie Vialaneix, and Pierre Neuvial. Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints. *Journal of Classification*, 38:363–389, 2021.

[13] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[14] Haoyue Zhang, Jessica Lam, Di Zhang, Yemin Lan, Marit W Vermunt, Cheryl A Keller, Belinda Giardine, Ross C Hardison, and Gerd A Blobel. Ctcf and transcription influence chromatin structure re-configuration after mitosis. *Nature communications*, 12(1):5157, 2021.

Données omiques et modèles mixtes

Regularized estimation of non-linear mixed effect models in the presence of high-dimensional biomarkers

Auriane Gabaut [1+], Ariane Bercu [2], Mélanie Prague [1*] & Cécile Proust-Lima [2*]

[1] *Université de Bordeaux, Inria, Inserm, Bordeaux Population Health Research Center ; Vaccine Research Institute, Créteil, France ;*
[2] *Université de Bordeaux, Inserm, Bordeaux Population Health Research Center, BIOSTAT Team*
∗ *co-last authors ;* + *auriane.gabaut@inria.fr*

Résumé. La construction de modèles non linéaires à effets mixtes (NLMEM) est précieuse pour enrichir notre compréhension des processus biologiques. Cependant, obtenir un modèle identifiable nécessite souvent l'observation de plusieurs compartiments, information que nous pensons pouvoir déduire à partir de données transcriptomiques. Ainsi, nous proposons une méthode de régularisation pour estimer des modèles non-linéaires mixtes impliquant des compartiments non observés, mesurés par des biomarqueurs en grande dimension, et sélectionner parmi eux les biomarqueurs pertinents.

Pour ce faire, nous utilisons un algorithme itératif d'estimation du maximum de vraisemblance pénalisée par Lasso (Tibshirani, 1996). L'algorithme d'estimation alterne entre une étape de régularisation et une étape d'inférence mécaniste, dans le but de maximiser la vraisemblance pénalisée du modèle. Lors de l'étape de régularisation, nous mettons à jour les coefficients associant le compartiment sous-jacent aux biomarqueurs en résolvant l'équation du Lasso, avec des approximations obtenues via un développement de Taylor d'ordre deux. L'étape d'inférence mécaniste se concentre sur l'estimation des paramètres mécanistes à l'aide de l'algorithme Stochastic Approximation Expectation-Maximization (SAEM) (Dempster *et al.*, 1977 ; Delyon *et al.*, 1999), implémenté via le logiciel Monolix (Kuhn et Lavielle, 2005 ; Lixoft SAS, a Simulations Plus company, 2023), en tenant compte des coefficients régularisés mis à jour lors de la première étape.

Nous présentons une étude de simulation illustrant les performances de notre méthode de sélection de variables et d'estimation des paramètres, ainsi qu'une application concrète utilisant des données de vaccins à ARNm contre la COVID-19 (Rinchai *et al.*, 2022).

Mots-clés. Apprentissage statistique ; Biostatistique ; génomique ; santé ; Grande dimension ; Sélection de modèles : Modèles non linéaires à effets mixtes.

Abstract. Mechanistic models are valuable for understanding the dynamic interactions between cell populations. However, achieving model identifiability often requires observations across multiple compartments, information that we believe could be inferred from transcriptomic data. Thus, we propose a regularization method to estimate nonlinear mixed-effects models involving unobserved compartments, measured through high-dimensional biomarkers, and to select the relevant biomarkers among them.

To achieve this, we employ an iterative algorithm for Lasso-penalized maximum likelihood estimation (Tibshirani, 1996).The estimation algorithm iterates between a regularization step and a mechanistic inference step with the overall goal to maximize the penalized likelihood of the model. In the regularization step, we update the coefficients associating the latent compartment to the biomarkers, derived from setting the first derivatives of the penalized log-likelihood to zero, with approximations obtained via second order Taylor developpement. The mechanistic inference step focuses on estimating the mechanistic parameters using the Stochastic Approximation Expectation-Maximization (SAEM) algorithm (Dempster *et al.*, 1977 ; Delyon *et al.*, 1999), implemented through the Monolix software (Kuhn et Lavielle, 2005 ; Lixoft SAS, a Simulations Plus company, 2023), considering the updated regularized coefficients from the first step.

We present a simulation study illustrating the performance of our proposed variable selection and parameter estimation method, along with a real-world application using COVID-19 mRNA vaccines data (Rinchai *et al.*, 2022).

Keywords. Statistical learning - Biostatistics, genomics, health - High dimension - Model selection

1. Introduction

Dans un contexte où les modèles mécanistes s'imposent pour analyser la dynamique des systèmes biologiques, l'intégration de biomarqueurs longitudinaux représente un enjeu majeur pour améliorer l'identifiabilité des paramètres. Pour répondre à cela, nous proposons une approche itérative basée sur l'estimation du maximum de vraisemblance pénalisée par Lasso, en recourant à un algorithme de descente cyclique par coordonnées (Friedman *et al.*, 2010 ; Tseng, 2001). Ces méthodes ont d'ores et déjà fait leurs preuves pour l'estimation de modèles conjoints intégrant des covariables en grande dimension dans les modèles de Cox (Caillebotte *et al.*, 2024), ainsi que pour la sélection et l'estimation d'effets fixes et aléatoires dans des modèles mécanistes (Pan et Huang, 2014 ; Li *et al.*, 2018 ; Yu et Xiang, 2024). Notre

objectif est d'étendre ces techniques afin d'estimer de manière robuste les paramètres liant les biomarqueurs longitudinaux aux compartiments sous-jacents dans un cadre mécaniste.

2. Méthode

2.1. Définition du modèle mécaniste

Supposons que nous ayons un système d'équations différentielles, appelé par la suite notre modèle structurel, contenant $P+1$ compartiments notés $(S_p)_{p\leq P}$ et R, qui dépendent de certains paramètres. Nous définissons alors un NLMEM en considérant les paramètres, notés $\psi = (\psi_l)_{l\leq m}$, du modèle structurel comme individuels à l'aide d'un modèle linéaire généralisé, $\forall l \leq m, i \leq N$:

$$h_l(\psi_{li}) = h_l(\psi_{lpop}) + X_i\beta_l + \eta_{li},$$

avec les paramètres de population $\psi_{pop} = (\psi_{lpop})_{l\leq m}$, les covariables de l'individu i $(X_i)_{i\leq N}$ associés aux paramètres $\beta = (\beta_l)_{l\leq m}$ et les effets aléatoires $\eta_i = (\eta_{li})_{l\leq m} \overset{iid}{\sim} \mathcal{N}(0, \Omega)$, où Ω est éventuellement creuse avec des termes diagonaux nuls,

Pour simplifier les formules, nous écrivons le processus individuel au cours du temps, résultant du système différentiel pour un ensemble de paramètres $\psi_i = (\psi_{li})_{l\leq m}$ pour l'individu $i \leq N$, sous la forme $S_p(\cdot, \psi_i) = S_{pi}(\cdot)$, $p \leq P$ et $R(\cdot, \psi_i) = R_i(\cdot)$.

Nous supposons que les trajectoires individuelles $(S_{pi})_{p\leq P, i\leq N}$ sont observées à travers un modèle d'observation direct, avec éventuellement une transformation fixée g_p, $p \leq P$, aux différents temps $(t_{pij})_{i\leq N, p\leq P, j\leq n_{ip}}$:

$$Y_{pij} = g_p(S_{pi}(t_{pij})) + \epsilon_{pij},$$

avec $\epsilon_p = (\epsilon_{pij}) \overset{iid}{\sim} \mathcal{N}(0, \Sigma_p^2)$ pour $p \leq P$.

Les trajectoires individuelles $(R_i)_{i\leq N}$ sont, quant à elles, observées à partir de K biomarqueurs $(Z_k)_{k\leq K}$ au travers de modèles d'observation linéaires généralisés :

$$Z_{kij} = \alpha_{k0} + \alpha_{k1} s_k(R_i(t_{kij})) + \varepsilon_{kij},$$

avec éventuellement une transformation s_k, $k \leq K$, les temps $(t_{kij})_{k\leq K, i\leq N, j\leq n_{kij}}$, et $\varepsilon_k \overset{iid}{\sim} \mathcal{N}(0, \sigma_k^2)$.

Les paramètres du modèle sont alors les paramètres de population du modèle mécaniste $\theta = (\psi_{pop}, \beta, \Omega, (\sigma_k^2)_{k\leq K}, (\Sigma_p^2)_{p\leq P}, (\alpha_{0k})_{k\leq K})$, et le vecteur de coefficients

liant les biomarqueurs en grande dimension au compartiment du modèle mécaniste sur lequel on souhaite appliquer une régularisation $\alpha = (\alpha_{1k})_{k\leq K} \in \mathbb{R}^K$, noté ultérieurement $\alpha = (\alpha_k)_{k\leq K} \in \mathbb{R}^K$ pour simplifier.

2.2. Procédure d'estimation

Nous proposons d'estimer ce type de modèle mécaniste avec régularisation sur les paramètres α par maximisation de la log-vraisemblance pénalisée.

2.2.1. Vraisemblance du modèle

Pour un individu $i \leq N$, la contribution à la vraisemblance est

$$\mathcal{L}_i(\theta, \alpha) = p\left(\left((Y_{pi})_{p\leq P}, (Z_{ki})_{k\leq K}\right)|\theta, \alpha\right) = \int_{\eta_i} f\left((Y_{pi})_{p\leq P}, (Z_{ki})_{k\leq K}|\theta, \alpha, \eta_i\right) \times f(\eta_i)d\eta_i$$

Conditionnellement aux paramètres et aux effets aléatoires, les trajectoires $(S_{pi})_{p\leq P}$ et R_i sont déterminées par le système différentiel, et ainsi les observations $(Z_{kij})_{k\leq K, j\leq n_{ki}}$ sont indépendantes et suivent une distribution $\mathcal{N}(\alpha_{0k}+\alpha_{k1}s_k(R_i(t_{kij})), \sigma_k^2)$. De même, $(Y_{pij})_{p\leq P, j\leq n_{pi}}$ suivent une distribution $\mathcal{N}(g_p(S_{pi}(t_{pij})), \Sigma_l^2)$. Nous sommes alors en mesure d'écrire la vraisemblance sous la forme :

$$\mathcal{L}_i(\theta, \alpha) = \int_{\eta_i} f(Z|\theta, \alpha, \eta_i) \times A(Y|\theta, \eta_i)d\eta_i$$

dont seul le premier facteur dans l'intégrande dépend des paramètres de régularisation α et des biomarqueurs, tandis que le second facteur dépend des dynamiques directement observées du reste des compartiments.

Enfin, les individus étant indépendants, nous pouvons définir la log-vraisemblance LL du modèle considéré comme :

$$LL(\theta, \alpha) = \sum_{i\leq N} LL_i(\theta, \alpha) = \sum_{i\leq N} \log\left(\int_{\eta_i} f(Z|\theta, \alpha, \eta_i) \times A(Y|\theta, \eta_i)d\eta_i\right)$$

2.2.2. Problèmes d'optimisation

Le problème que ce travail vise à résoudre est de trouver, pour $\lambda > 0$

$$(\theta_\lambda^*, \alpha_\lambda^*) = \underset{(\theta,\alpha)\in\Omega\times\mathbb{R}^K}{\operatorname{argmax}} LL_{pen}(\theta, \alpha) = \underset{(\theta,\alpha)\in\Omega\times\mathbb{R}^K}{\operatorname{argmax}} \left(LL(\theta, \alpha) - \lambda|\alpha|_1\right)$$

L'optimisation se fait de manière alternée sur les paramètres (θ, α) en utilisant une approche de descente de coordonnées cyclique. Dans un premier temps, pour un $\hat{\alpha} \in \mathbb{R}^K$ fixé, le problème revient à une maximisation classique de vraisemblance, qui peut être efficacement résolue par l'algorithme SAEM (Dempster *et al.*, 1977 ; Delyon *et al.*, 1999) dans le cadre des NLMEM (Kuhn et Lavielle, 2005).

Ensuite, pour un $\hat{\theta} \in \Omega$ fixé, il s'agit de résoudre $\left(\max_{\alpha} LL(\hat{\theta}, \alpha) - \lambda|\alpha|_1\right)$ qui revient à résoudre

$$\partial_{\alpha_{k'}} LL(\hat{\theta}, \alpha) - \lambda \operatorname{sign}(\alpha_{k'}) = 0.$$

Nous utilisons pour cela un développement de Taylor d'ordre 2 autour de $\hat{\alpha}$, qui correspond dans le cadre de notre méthode au paramètre α précédemment estimé :

$$LL(\hat{\theta}, \alpha) \approx LL(\hat{\theta}, \hat{\alpha}) + (\alpha - \hat{\alpha})^T \nabla_\alpha LL(\hat{\theta}, \hat{\alpha}) + \frac{1}{2}(\alpha - \hat{\alpha})^T H (\alpha - \hat{\alpha}),$$

avec $H = \nabla^2_\alpha LL(\hat{\theta}, \hat{\alpha}) - \hat{X} = -(\hat{x}_{lc})_{l,c \leq K}$ la matrice hessienne, supposée définie négative. L'équation devient alors

$$\partial_{\alpha_{k'}} LL(\hat{\theta}, \hat{\alpha}) - \sum_{l \neq k'} \hat{x}_{lk'} \alpha_l - \alpha_{k'} \hat{x}_{k'k'} + \sum_c \hat{x}_{k'c} \hat{\alpha}_c - \lambda \operatorname{sign}(\alpha_{k'}) = 0.$$

2.2.3. Implémentation

La méthode a été implémentée en `R` (R Core Team, 2024) avec les packages `lixoftConnectors` (LIXOFT, 2023) et `marqLevAlg` (Philipps *et al.*, 2021). Le code est disponible sur GitHub : `https://github.com/aurianegbt/REMix`. L'algorithme est présenté en figure 1.

Le calcul de la log-vraisemblance repose sur une quadrature de Gauss-Hermite multidimensionnelle, combinée avec un solveur permettant d'estimer simultanément R_i et $(S_{pi})_p$. Monolix est utilisé pour estimer la distribution a posteriori des effets aléatoires, afin d'adapter la quadrature, via le package `fastGHQuad` (Blocker, 2022).

Soit $\lambda > 0$ **fixé** ;
Soit $\theta^{(0)}, \alpha^{(0)}$, obtenus par exemple par MLE ou SAEM à partir d'un sous-ensemble aléatoire de biomarqueurs ;
On estime la distribution a posteriori des effets aléatoires par MCMC via Monolix
On calcule $LL\left(\theta^{(0)}, \alpha^{(0)}\right)$, $\partial_\alpha LL\left(\theta^{(0)}, \alpha\right)\Big|_{\alpha=\alpha^{(0)}}$ et $\partial^2_\alpha LL\left(\theta^{(0)}, \alpha\right)\Big|_{\alpha=\alpha^{(0)}}$;
A l'itération k **:**

On calcule $\alpha^{(k+1)}$ à partir de l'équation (**??**) avec $\hat{\alpha} = \alpha^{(k)}$ et $\hat{\theta} = \theta^{(k)}$ fixés ;
On estime $\theta^{(k+1)}$, avec $\alpha = \alpha^{(k+1)}$ fixé, via un algorithme SAEM partant de $\theta^{(k)}$;
On met à jour $LL\left(\theta^{(k+1)}, \alpha^{(k+1)}\right)$, $\partial_\alpha LL\left(\theta^{(k+1)}, \alpha\right)\Big|_{\alpha=\alpha^{(k+1)}}$ et $\partial^2_\alpha LL\left(\theta^{(k+1)}, \alpha\right)\Big|_{\alpha=\alpha^{(k+1)}}$ à partir d'une nouvelle estimation de la distribution a posteriori des effets aléatoires ;
On évalue les critères de convergence et arrête l'algorithme s'ils sont tous respectés :

- $\left|LL_{pen_\lambda}\left(\theta^{(k)}, \alpha^{(k)}\right) - LL_{pen_\lambda}\left(\theta^{(k-1)}, \alpha^{(k-1)}\right)\right| < \varepsilon_1$;
- $\left|\left(\theta^{(k)}, \alpha^{(k)}\right) - \left(\theta^{(k-1)}, \alpha^{(k-1)}\right)\right| < \varepsilon_2$.

Algorithm 1 : Algorithme REMix

Finalement, pour chaque valeur de λ utilisée, le modèle sélectionné correspond à celui minimisant le BICc. À cette étape, une dernière estimation par SAEM est effectuée sur le sous-ensemble de biomarqueurs Z sélectionnés pour éviter la contraction vers 0 des paramètres de régularisation, suivie de tests de Wald pour corriger les paramètres sélectionnés non significatifs (Wald, 1943).

3. Etude sur simulation

Nous avons simulé la production d'anticorps chez 25 individus, en considérant un compartiment de cellules sécrétrices d'anticorps (ASC) noté S, produisant des anticorps à un taux φ_S et décroissant à un taux δ_S. Les anticorps décroissent à un taux δ_{Ab}. Les paramètres individuels ont été simulés selon des lois log-normales incorporant des effets aléatoires.

Comme souvent observé dans les essais cliniques, et en se basant sur les essais cliniques des vaccins à ARNm contre la COVID-19 (Rinchai *et al.*, 2022), seules les concentrations d'anticorps aux jours 0, 7, 21, 123, 180 et 300 après vaccination sont observées. Le compartiment latent ASC est inféré à partir de biomarqueurs mesurés

quotidiennement entre les jours 1 et 21 après vaccination. Ces mesures comprennent $K_1 = 5$ biomarqueurs pertinents reflétant la dynamique des ASC, et $K_2 = 15$ biomarqueurs ne contenant que du bruit. Ce scénario a été répliqué 200 fois pour évaluer les performances de la méthode proposée.

Les performances en sélection de variables ont été évaluées à l'aide du taux de fausses découvertes (FDR), du taux de faux négatifs (FNR), définis comme suit :

$$FDR = \frac{FP}{FP + TP}$$

$$FNR = \frac{FN}{FN + TN}$$

ainsi que par le $F1$-score, qui équilibre précision et rappel :

$$F1\text{-score} = \frac{2TP}{2TP + FN + FP}$$

Cette métrique permet une évaluation globale du compromis entre faux positifs et faux négatifs.

La Figure 1 résume les performances en terme de sélection de variables. Dans notre scénario de simulation, nous observons de bonnes performance avec un FDR médian de 5.3% [0.0% ; 20.0%] et un FNR constant à 0%. Le $F1$-score médian atteint 97.1% [88.9% ; 100%]. Par ailleurs, les faux positifs sont dispersés parmi les biomarqueurs non pertinents, sans concentration particulière sur certains gènes. La fréquence moyenne de sélection des biomarqueurs non pertinents est faible, à 5.3%, et aucun biomarqueur ne présente un taux de sélection élevé, la fréquence maximale observée étant de 10.5%.

De plus, les résultats mettent en évidence la capacité de la méthode à estimer avec précision à la fois les paramètres populationnels (δ_{Ab}, δ_S, φ_S et leurs variances) et les paramètres spécifiques aux biomarqueurs (α_{0k}, α_{1k} et σ_k), avec des biais relatifs systématiquement inférieurs à 5% et des probabilités de couverture à 95% très proches du niveau nominal, avec une médiane de 95%.

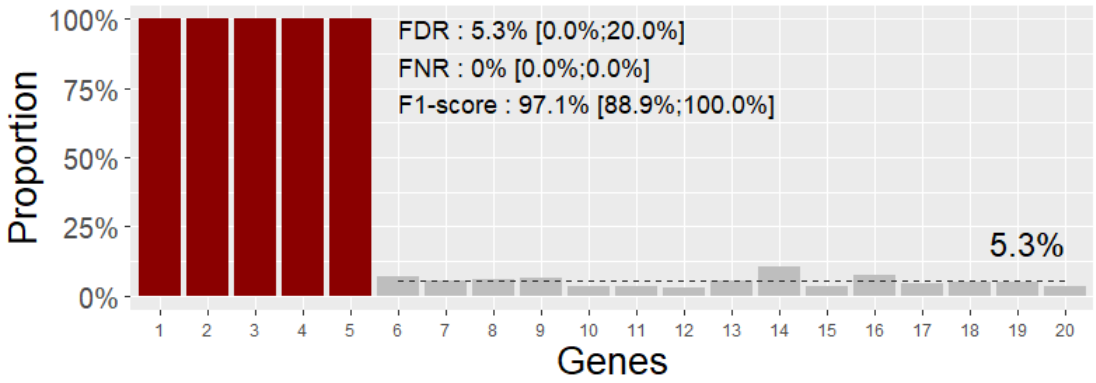


FIGURE 1 – Fréquence de sélection de chaque biomarqueur parmi 200 réplicats, avec les métriques de performance (FDR, FNR, $F1$-score) et leurs intervalles interquantiles à 95%.

4. Illustration

Nous évaluons les performances de notre algorithme en simulations. Puis, nous proposons une application de la méthode sur les données recueillies lors d'une étude sur la COVID-19 (Rinchai *et al.*, 2022). Cette étude vise à mesurer la réponse immunitaire après différents vaccins à ARN messager contre la COVID-19 et à mieux comprendre le lien entre transcriptomique et réponse immunitaire. Sur cette illustration, notre hypothèse de travail repose sur l'idée que l'expression des gènes constitue l'observation d'un processus latent décrivant les populations de cellules sécrétrices d'anticorps.

5. Discussion

Nous proposons une nouvelle approche d'estimation et de sélection dans des modèles mécanistes à effets mixtes, intégrant des biomarqueurs en grande dimension via une pénalisation de type Lasso. L'objectif est d'améliorer l'identifiabilité des paramètres associés à des compartiments non observés du modèle.

Nos simulations ont montré des performances prometteuses, caractérisées par un faible taux de fausses découvertes (FDR médian de 5.3%) et un taux de faux négatifs (FNR) nul, aboutissant à un excellent $F1$-score médian de 97.1%. Par ailleurs, les paramètres estimés démontrent une précision satisfaisante avec des biais relatifs

inférieurs à 5% et des probabilités de couverture proches du niveau nominal attendu de 95%. Des simulations complémentaires avec un plus grand nombre de gènes -50 gènes dont 10 informatifs- ont confirmé ces performances, montrant la robustesse de la méthode en plus grande dimension.

Ces résultats confirment l'intérêt de notre méthode pour identifier des biomarqueurs pertinents et améliorer l'interprétabilité des modèles mécanistes. Elle ouvre des perspectives en pharmacogénomique ou pharmacologie des systèmes. Des extensions possibles incluent la sélection de λ par des critères alternatifs et l'intégration de plusieurs processus latents.

Références

Blocker, A. W. (2022). *fastGHQuad : Fast 'Rcpp' Implementation of Gauss-Hermite Quadrature*. R package version 1.0.1.

Caillebotte, A., Kuhn, E. et Lemler, S. (2024). Estimation and variable selection in a joint model of survival times and longitudinal outcomes with random effects.

Delyon, B., Lavielle, M. et Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1) :94 – 128.

Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 39 :1–22.

Friedman, J. H., Hastie, T. et Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1) :1–22.

Kuhn, E. et Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4) :1020–1038.

Li, Y., Wang, S., Song, P. X., Wang, N., Zhou, L. et Zhu, J. (2018). Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data. *Statistics and its Interface*, 11 :721–737.

LIXOFT (2023). *lixoftConnectors : R connectors for Lixoft Suite (@Lixoft)*. R package version 2023.1.

Lixoft SAS, a Simulations Plus company (2023). Monolix. Version 2023R1.

Pan, J. et Huang, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Statistics and Computing*, 24 :725–738.

Philipps, V., Hejblum, Boris, P., Prague, M., Commenges, D. et Proust-Lima, C. (2021). Robust and efficient optimization using a marquardt-levenberg algorithm with r package marqlevalg. *The R Journal*, 13(2) :273.

R Core Team (2024). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rinchai, D., Deola, S., Zoppoli, G., Kabeer, B. S. A., Taleb, S., Pavlovski, I., Maacha, S., Gentilcore, G., Toufiq, M., Mathew, L., Liu, L., Vempalli, F. R., Mubarak, G., Lorenz, S., Sivieri, I., Cirmena, G., Dentone, C., Cuccarolo, P., Giacobbe, D. R., Baldi, F., Garbarino, A., Cigolini, B., Cremonesi, P., Bedognetti, M., Ballestrero, A., Bassetti, M., Hejblum, B. P., Augustine, T., Panhuys, N. V., Thiebaut, R., Branco, R., Chew, T., Shojaei, M., Short, K., Feng, C. G., Consortium, P.-., Zughaier, S. M., Maria, A. D., Tang, B., Hssain, A. A., Bedognetti, D., Grivel, J.-C. et Chaussabel, D. (2022). High–temporal resolution profiling reveals distinct immune trajectories following the first and second doses of covid-19 mrna vaccines. *Science Advances*, 8(45) :eabp9961.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58 :267–288.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization 1.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54 :426–482.

Yu, T. et Xiang, L. (2024). An iterative two-step regularization approach for selection of fixed and random effects in generalized linear mixed models. *Japanese Journal of Statistics and Data Science*.

phoenics : UNE NOUVELLE APPROCHE STATISTIQUE POUR L'ANALYSE DIFFÉRENTIELLE LONGITUDINALE DE VOIES MÉTABOLIQUES

Camille Guilmineau [1], Rémi Servien [1], Marie Tremblay-Franco [2,3], Nathalie Vialaneix [4]

[1] *INRAE, Université de Montpellier, LBE, F-11100, Narbonne, France.*
{*camille.guilmineau, remi.servien*}*@inrae.fr*
[2] *INRAE, Université de Toulouse, ENVT, Toxalim, Toulouse F-31027, France.*
[3] *Axiom Platform, MetaToul-MetaboHUB, National Infrastructure for Metabolomics and Fluxomics, Toulouse F-31027, France.*
marie.tremblay-franco@inrae.fr
[4] *Université de Toulouse, INRAE, UR MIAT, Castanet-Tolosan F-31326, France.*
nathalie.vialaneix@inrae.fr

Résumé. La métabolomique décrit le profil métabolique d'un organisme par les concentrations des métabolites qui le compose à un instant donné. Lorsqu'elles sont étudiées au fil du temps, les concentrations des métabolites peuvent aider à comprendre l'évolution dynamique d'un processus biologique. Cependant, ces métabolites sont impliqués dans des séquences de réactions chimiques, appelées voies métaboliques, liées à une fonction biologique donnée. La prise en compte de ces voies dans les méthodes statistiques d'analyse des données métabolomiques est donc un moyen pertinent d'exprimer directement les résultats en termes de fonctions biologiques et d'augmenter leur interprétabilité. Nous proposons une nouvelle méthode, **phoenics**, qui effectue une analyse différentielle des données métabolomiques longitudinales au niveau des voies. En bref, **phoenics** procède en deux étapes : tout d'abord, la matrice de quantification des métabolites est transformée par une approche de réduction de dimension prenant en compte les informations sur les voies. Ensuite, un modèle linéaire mixte est ajusté sur les données transformées. Cette méthode a été appliquée à des données semi-synthétiques et à un jeu de données réelles évaluant les effets des antibiotiques sur les matières fécales. Les résultats ont montré que **phoenics** contrôle correctement le taux d'erreur de type I et a une meilleure capacité à détecter les voies métaboliques différentielles et à extraire de nouvelles fonctions biologiques impactées que les méthodes alternatives. La méthode est implémentée dans le package R **phoenics** disponible sur CRAN.

Mots-clés. modèles à effets mixtes, données longitudinales, métabolomique, voies métaboliques.

Abstract. Metabolomics describes the metabolic profile of an organism at a given time by the concentrations of its constituent metabolites. When studied over time, metabolite concentrations can help understand the dynamical evolution of a biological process. However, metabolites are involved into sequences of chemical reactions, called metabolic pathways, related to a given biological function. Accounting for these pathways into statistical methods for metabolomic data is thus a relevant way to directly express results in terms of biological functions and to increase their interpretability. We propose a new method, **phoenics**, to

perform differential analysis for longitudinal metabolomic data at the pathway level. In short, **phoenics** proceeds in two steps: First, the matrix of metabolite quantifications is transformed by a dimension reduction approach accounting for pathway information. Then, a mixed linear model is fitted on the transformed data. This method was applied to semi-synthetic data and a real dataset assessing effects of antibiotic on feces. Results showed that **phoenics** properly controls the Type I error rate and has a better ability to detect differential metabolic pathways and to extract new impacted biological functions than alternative methods. The method is implemented in the R package **phoenics** available on CRAN.

Keywords. mixed model, longitudinal data, metabolomics, metabolic pathways.

1 Introduction

Les données métabolomiques correspondent à la quantification de la concentration de petites molécules, appelées métabolites, présentes dans des mélanges complexes à un instant donné. De plus en plus d'expériences ciblent l'évolution du métabolome au cours du temps, dans différentes conditions, ou suite à un événement d'intérêt (Mäkinen and al., 2022). Dans ces cas, les données métabolomiques sont acquises à plusieurs moments sur les mêmes individus. Nous abordons ici la question de l'utilisation de données métabolomiques temporelles acquises dans plusieurs conditions expérimentales pour extraire des biomarqueurs soit entre différentes conditions, soit entre différents temps de mesure.

Les métabolites sont cependant regroupés en séquences de réactions chimiques, appelées voies métaboliques, dans lesquelles ils sont impliqués et qui partagent une fonction biologique commune. La prise en compte des voies dans les méthodes statistiques pour les données métabolomiques est donc un bon moyen d'exprimer directement les résultats en termes de fonctions biologiques et d'augmenter leur interprétabilité. Habituellement, cela se fait par analyse d'enrichissement, une méthode qui consiste à effectuer des tests indépendamment pour chaque métabolite et à combiner les résultats de ces tests pour extraire des voies contenant plus de métabolites différentiels que ce qui aurait été attendu par hasard (Khatri and al., 2012). Un inconvénient est que cette approche est basée sur la puissance d'une analyse primaire où les métabolites sont analysés de manière indépendante est donc incapable de détecter des voies qui correspondent a une accumulation de petits effets.

D'autres méthodes, dites de notation de classe fonctionnelle (FCS), ont également été utilisées pour effectuer des analyses différentielles. La plupart de ces méthodes ont été développées dans le domaine de la transcriptomique mais elles peuvent être facilement étendues à l'analyse des voies métabolomiques (Maleki and al., 2020). Dans le domaine de la métabolomique, Wieder and al. (2022, 2024) développent une de ces méthodes, où les données au niveau du métabolite sont transformées en données au niveau de la voie avant qu'une analyse différentielle soit réalisée sur cette transformation. Aussi, Ozier-Lafontaine and al. (2024) proposent un test non linéaire pour les ensembles de gènes basé sur une approche par noyau. Dans son implémentation actuellement disponible, le test est cependant limité aux différences de test entre deux conditions sans covariables.

Nous présentons ici une extension de la méthode de Wieder and al. (2022) pour per-

mettre l'analyse de mesures longitudinales répétées. Notre méthode, **phoenics**, procède en deux étapes : une transformation des données similaire à Wieder and al. (2022) et un modèle mixte. La méthode est implémentée dans le package R **phoenics** disponible sur le CRAN[1].

2 Méthode

Soit X une matrice $(nT \times m)$ de quantifications, où n est le nombre d'individus, T le nombre de points dans le temps et m le nombre de métabolites. X est organisée de telle sorte que ses n premières lignes correspondent aux mesures de quantification des n individus au premier point dans le temps. Chaque métabolite appartient à une ou plusieurs voies métaboliques $\mathcal{M}_l$. La méthode proposée est composée de deux étapes :

1. **Représentation par analyse factorielle pour extraire les signatures des voies.** Pour chaque voie métabolique $\mathcal{M}_l$, et de manière similaire à Wieder and al. (2022), une ACP est réalisée sur la matrice de quantification des métabolites de cette voie, $Z_l = (X_{ij})_{i=1,\dots,nT,\, j\in\mathcal{M}_l}$. Les scores des m_l^* premières composantes principales sont extraits dans la $(nT \times m_l^*)$-matrice A_l. Dans ce qui suit, pour simplifier les notations, nous ferons référence génériquement à a comme l'une des colonnes d'une des matrices A_l, qui correspond à l'une des composantes principales pour une voie donnée.

2. **Modèle mixte pour tester les voies.** Pour chaque composante principale a, le modèle mixte suivant est alors estimé :

$$a = \beta_t + \theta_d + u_i + \epsilon_{tdi} \tag{1}$$

avec

- β_t l'effet fixe du temps t ;
- θ_d l'effet fixe de la condition d ;
- $u_i \sim \mathcal{N}(0, G)$ l'effet aléatoire de l'individu i, avec G la matrice de covariance ;
- $\epsilon_{tdi} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ les résidus du modèle, qui sont supposés être i.i.d.

La signification d'un effet fixe donné f est ensuite testée à l'aide du test du rapport de vraisemblance comparant le modèle complet de l'équation (1) à un modèle restreint qui ne contient pas l'effet f.

Pour une voie métabolique donnée $\mathcal{M}_l$, cette estimation est répétée pour chaque colonne de A_l, ce qui donne m_l^* p-valeurs. La procédure de Simes (1986) est utilisée pour agréger les m_l^* p-valeurs : cette procédure contrôle l'erreur de type I de l'hypothèse nulle $\mathrm{H}_0 = \cap_{j=1}^{m_l^*} \mathrm{H}_{0j}$, où H_{0j} est l'égalité des moyennes dans les différents niveaux de f pour la $j$ième colonne de A_l. En conséquence, cette approche nous conduit à obtenir une seule valeur p par voie métabolique pour chaque effet testé f. Enfin, comme plusieurs valeurs p sont obtenues (une pour chaque voie), une correction supplémentaire par test multiple est nécessaire. Nous avons choisi de contrôler le FDR en utilisant la procédure de Benjamini-Hochberg (BH).

[1] `https://CRAN.R-project.org/package=phoenics`

En pratique, la méthode nécessite seulement de choisir le paramètre m_l^*, correspondant au nombre de composantes principales sur lesquelles les modèles mixtes sont estimés pour chaque voie. Pour permettre à la méthode de conserver suffisamment d'informations dans les données pour estimer chaque effet possible, nous proposons de fixer m_l^* égal au nombre d'effets fixes, F.

Une limitation de cette approche est que l'ACP ne tient pas compte de la structure spécifique des données où certaines observations d'un individu donné peuvent être plus corrélées que d'autres. Ainsi, nous proposons également de remplacer l'ACP globale par des analyses ACP indépendantes, dans le même esprit que l'analyse multi-tables appelée Analyse Factorielle Multiple (MFA) (Escofier and Pagès, 1994). La matrice de quantification des métabolites de la voie testée est alors vue comme un ensemble de matrices, chacune correspondant à un instant donné, $t : Z_l^{(t)} = (X_{ij})_{i=n(t-1)+1,\ldots,nt,\, j\in\mathcal{M}_l}$ pour $t = 1, \ldots, T$. La MFA effectue ensuite une ACP individuelle (dite partielle) de chaque matrice (une ACP globale est aussi disponible, qui correspond à une moyenne pondérée des ACP individuelles). Ici, nous nous en tenons à la première étape de la MFA et récupérons les scores des m_l^* premières composantes principales des analyses partielles pour les empiler dans une $(nT \times m_l^*)$-matrice A_l. La deuxième étape (correspondant au modèle mixte) est ensuite réalisée de la même manière que l'approche précédente. Cependant, si cette approche paraît plus appropriée pour tester l'effet condition, elle n'est pas pertinente pour l'effet temps, car ce dernier devrait être effacé en effectuant les analyses partielles indépendantes.

3 Résultats

Données : Les performances de la méthode ont été évaluées à l'aide de données expérimentales de Choo and al. (2017), disponibles dans l'entrepôt de données MetaboLights avec l'identifiant MTBLS422. Cette étude examine les changements du métabolisme fécal induits par les antibiotiques. Les données contiennent des spectres de résonance magnétique nucléaire (RMN) obtenus à partir de 2×8 fluides de souris soumises ou non à un traitement antibiotique. Dans la suite, le statut de traitement de la souris est appelé la « condition » et est un effet fixe d'intérêt. De plus, dans chaque condition, des spectres ont été acquis à partir d'échantillons collectés à trois moments et l'effet fixe temporel est donc un deuxième effet fixe d'intérêt.

Les quantifications des métabolites ont été estimées à partir des spectres RMN avec le package R **ASICS** (Lefort and al., 2019). Cela a permis de quantifier 176 métabolites au total. Les voies métaboliques ont été récupérées à l'aide du package R **KEGGREST** qui interroge la base de données KEGG pour les voies spécifiques à un organisme donné (ici, *Mus musculus*), comme implémenté dans notre package **phoenics**. Un nombre total de 98 voies ont été obtenues, contenant au moins un des métabolites quantifiés pour un total de 128 métabolites trouvés dans les voies.

Ce jeu de données a ensuite été utilisé pour générer des jeux de données semi-synthétiques permettant de tester un effet temps, un effet condition et les deux effets simultanément. Ceci avec différentes intensités de l'effet, d'un effet fort (Scenario 1) à un effet faible (Scenario 4).

Les jeux de données semi-synthétiques ont permis de tester 3 caractéristiques des méthodes : le contrôle de l'erreur de type I, la puissance statistique et une analyse de la puissance selon le nombre de métabolites différentiels dans voie testée.

Méthodes : Nous avons comparé **phoenics** à la méthode de référence pour l'analyse des voies métaboliques, qui est l'analyse d'enrichissement. Pour cette approche, nous avons utilisé le même modèle mixte linéaire que dans **phoenics** mais par métabolite et non par voie et avons également corrigé les p-valeurs issues du test de ce modèle par la méthode BH. **phoenics** a aussi été comparé à une méthode FCS multivariée, **ktest**. Comme sa mise en œuvre est actuellement limitée à deux conditions, des comparaisons par paires des trois dates ont été réalisées pour le test de l'effet du temps, en utilisant les deux premières dimensions de l'espace image associé au noyau. Les deux premières dimensions ont également été utilisées pour le test de l'effet de condition. De manière similaire à ce qui est fait dans **phoenics**, la procédure de Simes a été utilisée pour agréger les six (2 dimensions $\times 3$ paires de dates) ou les deux (2 dimensions $\times 1$ paire de conditions) valeurs p résultantes. La multiplicité des tests a été prise en compte par la méthode BH. Dans la suite, phoenicsPCA fera référence à l'approche basée sur l'ACP alors que l'approche basée sur la MFA sera notée phoenicsMFA.

Résultats : Le contrôle de l'erreur de Type I a toujours été assuré pour toutes les méthodes. Pour le jeu de données incluant seulement un effet condition, phoenicsMFA s'est avéré plus puissant que phoenicsPCA et que les autres méthodes, les différences de performances se creusant quand l'amplitude de l'effet diminue. Cependant, comme supposé, lorsqu'il y a seulement un effet temps phoenicsMFA (tout comme **ktest**) ne détecte pas cet effet. phoenicPCA est alors plus performant que l'analyse d'enrichissement et la différence de performances est également plus forte quand l'effet est faible.

Pour le jeu de données comprenant les 2 effets, le test de l'effet temps a amené aux mêmes conclusions que précédemment. Le test de l'effet condition est plus intéressant, phoenicsPCA a une sensibilité plus élevée que phoenicsMFA et que l'analyse d'enrichissement dans tous les scénarios (Figure 1). De plus, phoenicsPCA présente un nombre plus faible de faux positifs par rapport à phoenicsMFA et a donc une Valeur Prédictive Positive (VPP) plus élevée. Aussi, **ktest** présente une sensibilité et une VPP élevées dans le scénario 1 mais ne détecte plus de positif lorsque la taille de l'effet diminue dans les autres scénarios.

L'application de **phoenics** sur le jeu de données réelles a ensuite permis de détecter les voies pertinentes pour l'effet des antibiotiques sur la métabolomique des matières fécales.

4 Conclusion

Nous présentons ici une approche permettant de réaliser une analyse différentielle longitudinale des voies métaboliques, disponible dans le package R **phoenics**. La méthode est basée sur l'analyse factorielle et des modèles linéaires mixtes pour réaliser l'analyse au niveau des voies et ne nécessite pas d'analyse préalable des métabolites. L'application sur des données a montré que les méthodes **phoenics** détectent les voies différentielles avec une puissance supérieure aux méthodes existantes tout en contrôlant le taux d'erreur de type I.

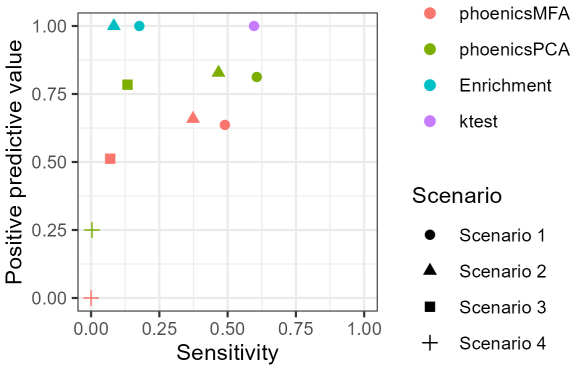


FIGURE 1 : VPP et sensibilité pour tester l'effet condition. L'analyse d'enrichissement pour les Scenarios 3 et 4 et **ktest** pour les Scenarios 2, 3 et 4 n'apparaissent pas car ils ne font pas de détections positives.

Cette recherche a été financée par l'Agence Nationale de la Recherche (ANR) au titre du projet ANR-21-CE45-0036-01.

References

Choo, J. M. and al. (2017). Divergent relationships between fecal microbiota and metabolome following distinct antibiotic-induced disruptions. *mSphere*, 2(1).

Escofier, B. and Pagès, J. (1994). Multiple factor analysis (AFMULT package). *Computational Statistics and Data Analysis*, 18(1):121–140.

Khatri, P. and al. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2):e1002375.

Lefort, G. and al. (2019). ASICS: an R package for a whole analysis workflow of 1D 1H NMR spectra. *Bioinformatics*, 35(21):4356–4363.

Maleki, F. and al. (2020). Gene set analysis: challenges, opportunities and future research. *Frontiers in Genetics*, 11:654.

Mäkinen, V.-P. and al. (2022). Longitudinal profiling of metabolic ageing trends in two population cohorts of young adults. *International Journal of Epidemiology*, 51:1970–1983.

Ozier-Lafontaine, A. and al. (2024). Kernel-based testing for single-cell differential analysis. *Genome Biology*, 25:114.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Wieder, C. and al. (2022). Single sample pathway analysis in metabolomics: performance evaluation and application. *BMC Bioinformatics*, 23:481.

Wieder, C. and al. (2024). PathIntegrate: multivariate modelling approaches for pathway-based multi-omics data integration. *PLoS Computational Biology*, 20(3):e1011814.

Modèle animal multi-caractères à copule pour améliorer la sélection.

Tom Rohmer [1], Victoria Brüning [1], Estelle Kuhn [2]

[1] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan*
[2] *Université Paris-Saclay, INRAE, AgroParisTech, MaIAGE, Jouy-en-Josas*
contact : tom.rohmer@inrae.fr.

Résumé.
En génétique animale, les modèles linéaires mixtes sont utilisés pour séparer les effets génétiques des effets environnementaux (modèle G+E). Les composantes de variance de ces modèles sont souvent estimées par maximum de vraisemblance restreint (REML), en supposant la normalité jointe du modèle d'inférence. Cependant, en pratique, lorsqu'on étudie simultanément plusieurs caractères, l'hypothèse de normalité multivariée pour les phénotypes peut être violée, notamment en raison d'une structure de dépendance non-normale entre les phénotypes. Un article récent, établit via une étude de simulation, qu'utiliser un modèle animal multi-caractères G+E Gaussien pour des caractères échantillonnés au travers une copule non normale pour la partie résiduelle peut introduire un biais dans l'estimation des paramètres génétiques, tels que l'héritabilité ou les corrélations génétiques, en présence de populations sous sélection non aléatoire des reproducteurs.

Pour résoudre ce problème, nous proposons un modèle d'inférence plus générique, non nécessairement Gaussien, qui prend en compte non seulement les parties génétique et environnementale, mais aussi une copule sur les termes résiduels, pouvant être non normale. Nous avons développé une stratégie de descente de gradient stochastique pour estimer conjointement les composantes de la variance (génétique et résiduelle) ainsi que les paramètres de la copule par maximum de vraisemblance. Nous comparerons les estimations des paramètres génétiques et résiduels obtenues à partir du modèle d'inférence non gaussien avec celles du modèle classique G+E Gaussien multi-caractères, estimés par approche REML. Cette comparaison utilisera des données simulées à partir du modèle d'inférence à copule de Clayton.
Mots-clés. Génétique quantitative, modèle mixte multivarié, copule.

Abstract.
In animal genetics, linear mixed models are used to address genetic and environmental effects (referred as G+E model). Variance components are often estimated using the restricted maximum likelihood (REML) approach, using normality assumptions. However, in practice, when dealing with multiple traits, the assumption of multivariate normality for the phenotypes may be violated, particularly due to a non-normal structure between the phenotypes. This is notably caused by a non-Normal copula in the environmental part,

i.e. residuals. It is established in a recent paper that using a Gaussian multitrait animal model (G+E) for traits sampled through a non-Normal copula for the residual part can bias the estimated genetic parameters, such as heritability or genetic correlations, especially in population undergoing non-random selection of reproducers. To address this issue, we propose a non-Gaussian inference model that considers not only the genetic and environmental parts, but also a copula on the residual terms, which can be non-Normal. We developed stochastic gradient strategies to maximize the considered log-likelihoods, allowing us to jointly estimate the variance components (genetic and residual) and copula parameters. We will compare the estimations of the genetic and residual parameters obtained from the non-Gaussian inference model with those from classical G+E Gaussian multitrait models, estimated by 'Average Information' -REML. This comparison will use simulated data generated from the Clayton copula inference model.

Keywords. Quantitative genetic, mixed model, multitrait model, copula.

1 Introduction

En génétique animale, les modèles mixtes sont fréquemment utilisés afin de dissocier la part de variabilité due à la génétique de la part de variabilité due à l'environnement sur les phénotypes étudiés, Mrode (2014). Ces derniers sont étendus au cas bivarié pour tenir compte des covariances entre caractères lorsque deux phénotypes $(y_{i,1}, y_{i,2})$ sont observés pour l'animal $i = 1, \ldots, n$ voir par ex. Meyer (1985) :

$$\left\{ \begin{array}{ll} \boldsymbol{y}_1 & = X_1\boldsymbol{\beta}_1 + Z\boldsymbol{a}_1 + \boldsymbol{\varepsilon_1} \\ \boldsymbol{y}_2 & = X_2\boldsymbol{\beta}_2 + Z\boldsymbol{a}_2 + \boldsymbol{\varepsilon_2}, \end{array} \right. \tag{1}$$

où pour $j = 1, 2$, $\boldsymbol{y}_j = (y_{1,j}, \ldots, y_{n,j})$, sont les caractères observées (i.e. les observations), X_j et Z les matrices d'incidences reliées aux effets fixes $\boldsymbol{\beta}_j$ et aux valeurs génétiques 'breeding values' (BV) $\boldsymbol{a}_j = (a_{1,j}, \ldots, a_{N,j})$, $N \geq n$, non-observables, pouvant être interprétées comme des variables latentes dans le modèle d'inférence. Enfin $\boldsymbol{\varepsilon_j} = (\varepsilon_{1,j}, \ldots, \varepsilon_{n,j})$ sont les termes résiduels. Les BV sont supposés Gaussiens, de matrice de variance covariance

$$var(\boldsymbol{a}_1, \boldsymbol{a}_2) = G \otimes A,$$

où A est une matrice de relation de parenté ('kinship') connue, de dimension $N \times N$, reliée au pedigree. La matrice G est la matrice de variance-covariance de dimension 2×2, inconnue à estimer. Enfin '$\otimes$' désigne le produit de Kronecker. Les résidus du modèle $(\epsilon_{i,1}, \epsilon_{i,2})$ sont supposés indépendants pour $i = 1, \ldots, n$ et indépendants des BV, $\epsilon_{i,j}$ sont Gaussiens, $j = 1, 2$, de variance σ_j^2 à estimer et $(\epsilon_{i,1}, \epsilon_{i,2})$ admettent pour copule une copule C_θ dont le paramètre θ est inconnu à estimer.

Plus précisément, considérons un vecteur aléatoire $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \mathcal{X}_2)$ de fonction de répartition $\boldsymbol{F}$, et de fonctions de répartitions marginales F_1, F_2 supposées continues. Sklar (1959) permet d'affirmer qu'il existe une unique fonction $C : [0,1]^2 \to [0,1]$ telle que :

$$\boldsymbol{F}(\boldsymbol{x}) = C\{F_1(x_1), F_2(x_2)\}, \quad \boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2. \tag{2}$$

Cette fonction C dite copule caractérise la structure de dépendance du vecteur aléatoire $\boldsymbol{\mathcal{X}}$. En particulier une structure de dépendance non-Gaussienne entraîne une distribution bivariée non-gaussienne pour les résidus.

Les composantes de variances sont généralement estimées par des méthodes de maximum de vraisemblance restreinte (REML) sous l'hypothèse de normalité jointe des phénotypes bien qu'en pratique cette hypothèse n'est pas toujours réaliste, en particulier de part des résidus non nécessairement Gaussiens.

Rohmer et al. (2022) ont établi qu'utiliser un modèle animal multi-caractères Gaussien (G+E) pour des caractères échantillonnés au travers d'une copule non normale pour la partie résiduelle peut introduire un biais dans l'estimation des paramètres génétiques, tels que l'héritabilité des caractères ou les corrélations génétiques, en présence de populations sous sélection non-aléatoire des reproducteurs.

La vraisemblance pour les valeurs génétiques est

$$\mathcal{L}_a(G; \boldsymbol{a}) = \frac{1}{(2\pi)^n (det(G \otimes A))^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{a}^T G^{-1} \otimes A^{-1}\boldsymbol{a}\right).$$

En utilisant le théorème de Sklar (2), en notant le vecteur des paramètres marginaux $\boldsymbol{\alpha} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2)$, la vraisemblance portée par les résidus $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2)$ est

$$\mathcal{L}_{y|a}(\boldsymbol{\alpha}, \theta; \boldsymbol{\varepsilon}) = \prod_{i=1}^{n} c_\theta\left(F_1(\varepsilon_{i,1}, \sigma_1^2), F_2(\varepsilon_{i,2}, \sigma_2^2)\right) \times \prod_{i=1}^{n}\prod_{j=1}^{2} f_j(\varepsilon_{i,j}, \sigma_j^2).$$

Ici F_j et f_j sont les fonctions de répartition et densités des résidus $\varepsilon_{i,j} = y_{i,j} - (\boldsymbol{z}_i^T a_j + \boldsymbol{x}_{ij}\boldsymbol{\beta}_j)$ gaussiens, de paramètres de variance σ_j^2. La fonction c_θ désigne la densité de la copule C_θ.

Afin d'estimer le paramètre joint $\boldsymbol{\eta} = (\boldsymbol{\alpha}, \theta, G)$, nous nous intéressons ici à une approche par maximum de vraisemblance (MLE) du modèle joint. Notons que la log-vraisemblance du modèle statistique s'écrit :

$$\log \mathcal{L}_{\boldsymbol{y}}(\boldsymbol{\alpha}, \theta, G; \boldsymbol{y}) = \log \int_{2N} \mathcal{L}_{y|a}(\boldsymbol{\alpha}, \theta; \boldsymbol{\varepsilon}) \mathcal{L}_a(G; \boldsymbol{a}) d\boldsymbol{a}. \tag{3}$$

Excepté pour des modèles typiques (par.ex. copule normale et distributions marginales gaussiennes), la log-vraisemblance $\log \mathcal{L}_{\boldsymbol{y}}$ dans (3) ne possède pas de forme explicite. Ainsi le MLE

$$\hat{\boldsymbol{\eta}} = \text{argmax}_{(\boldsymbol{\alpha}, \theta, G)} \log \mathcal{L}_{\boldsymbol{y}}(\boldsymbol{\alpha}, \theta, G; \boldsymbol{y})$$

pour le paramètre vectoriel $\boldsymbol{\eta} = (\boldsymbol{\alpha}, \theta, G)$ peut être difficile à atteindre.

Néanmoins nous cherchons à résoudre l'équation $\nabla_\eta \log \mathcal{L}_{\boldsymbol{y}}(\eta; \boldsymbol{y}) = 0$. Or d'après l'identité de Fisher (1970),

$$\nabla_\eta \log \mathcal{L}_{\boldsymbol{y}}(\boldsymbol{\eta}; \boldsymbol{y}) = \nabla_\eta \log \mathbf{E}_{\boldsymbol{a}} \left(\mathcal{L}_{y|a}(\boldsymbol{\eta}; \boldsymbol{y}|\boldsymbol{a}) \right) = \mathbf{E}_{\boldsymbol{a}|\boldsymbol{y}} \left(\nabla_\eta \log \mathcal{L}_{(\boldsymbol{y},\boldsymbol{a})}(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{a}) \right), \quad \text{avec } \boldsymbol{\eta} = (\boldsymbol{\alpha}, \theta, G),$$

où $\mathcal{L}_{(\boldsymbol{y},\boldsymbol{a})}$ représente la vraisemblance de la distribution jointe $(\boldsymbol{y}, \boldsymbol{a})$. Nous proposons ici une approche de descente de gradient stochastique (SGD) pour atteindre le MLE. L'algorithme est le suivant :

Initialization : $\boldsymbol{\eta}^{(0)}$;
for $m \in 1, \ldots, maxit$ **do**
1. Simuler $\boldsymbol{a}^{(m-1)}$ à partir de la distribution conditionnelle $\boldsymbol{a}|\boldsymbol{Y}$;
2. Mises à jour du paramètre :

$$\boldsymbol{\eta}^{(m)} = \boldsymbol{\eta}^{(m-1)} + \gamma_m \nabla_\eta \log \mathcal{L}_{(\boldsymbol{y},\boldsymbol{a})}(\boldsymbol{\eta}^{(m-1)}; \boldsymbol{y}, \boldsymbol{a}^{(m-1)}) \tag{4}$$

pour un pas d'apprentissage γ_m bien choisi.
end

Algorithm 1: SGD algorithm

Pour la copule non-Normale, le point 1. peut être compliqué et une approche par MCMC Métropolis-within-gibbs Robert et al. (1999) est mise en place. De par la parcimonie de A^{-1}, le point .2 n'est quant-à lui pas une limitation d'un point de vue numérique.

2 Résultats de simulations

Les populations d'animaux apparentés ont été générées en suivant Rohmer et al. (2022), mimant un schéma de sélection pouvant être rencontré dans un élevage porcin. Les fondateurs (G_0) consistaient en 102 femelles et 6 mâles non apparentés. Dans les générations suivantes (G_1–G_8), chaque mâle reproducteur a été accouplé à 17 femelles pour produire 10 femelles et 2 mâles phénotypés, soit $n = 9900$ individus à l'issu de 8 générations. Dans les 3 premières générations, les mâles et femelles reproducteurs ont été choisi au hasard. Dans les 5 générations suivantes, les valeurs génétiques des animaux pour les deux caractères ont été prédites par BLUP voir Henderson (1953) pour l'approche REML, où en utilisant l'échantillon issu de la distribution conditionnelle obtenue par MCMC pour l'approche SGD. Au sein de chaque famille de même père, les individus dont la somme de leurs BV prédits étaient les plus fortes ont été sélectionné (sélection par troncation). Dans chaque génération, un ratio de sélection pour la reproduction de 2.9% de mâles et 10% des femelles a été considéré. Les accouplements ont été réalisés au hasard de telle sorte que les pleins frères et demi-frères ne soient jamais accouplés afin de

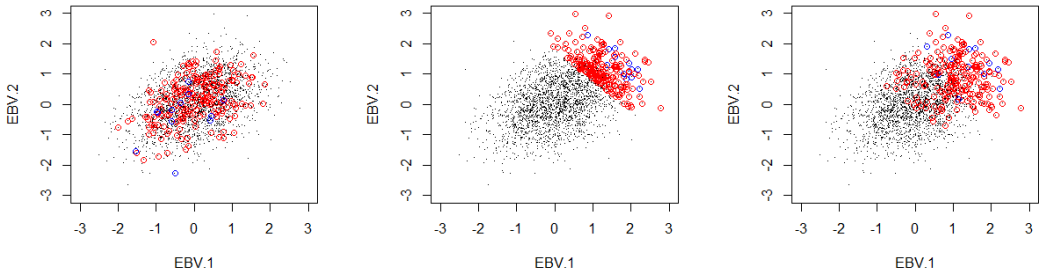


FIGURE 1 – selection : (left) at random (middle) by truncation (right) by intra-family truncation ; (red points) selected females (blue points) selected males (black points) non-selected animals

réduire la consanguinité. Le processus de sélection des reproducteurs est illustré dans la Figure 1.

Les vecteurs des résidus du modèle mixte ont été échantillonnés indépendamment selon une distribution bivariée avec des marges Gaussiennes de variance σ_j^2 fixée à 1 et de copule Clayton avec corrélation de Kendall fixée à 0.7, voir Nelsen (2007). La corrélation génétique théorique a été fixée à 0.31. Les variances génétiques théoriques étaient de 0.67.

De la génération G_3 à la génération G_8, nous avons estimé l'héritabilité des caractères ainsi que les corrélations génétiques et résiduelles par REML sous l'hypothèse de normalité jointe des phénotypes (c.-à-d. copule Normale) ainsi qu'en utilisant l'approche SGD en utilisant le modèle d'inférence à copule pour la copule de Clayton. Les pas d'apprentissages ont été calibrés en utilisant une approche adaptative de type adagrad.

Comme dénoté dans Rohmer et al. (2022), nous retrouvons dans la Figure 2 que l'utilisation d'un modèle classique Gaussien va engendrer un biais d'estimation en présence de sélection des reproducteurs (G_4-G_8) pour des données issu d'un modèle de type Clayton. L'inférence dans un modèle plus approprié, prenant en compte la structure de dépendance des données permet de corriger ces biais et de fournir des estimateurs des composantes de variance plus adéquats.

Références

Fisher, R. A. (1970), Statistical methods for research workers, *in* 'Breakthroughs in statistics : Methodology and distribution', Springer, pp. 66–70.

Henderson, C. R. (1953), 'Estimation of variance and covariance components', *Biometrics* **9**(2), 226–252.

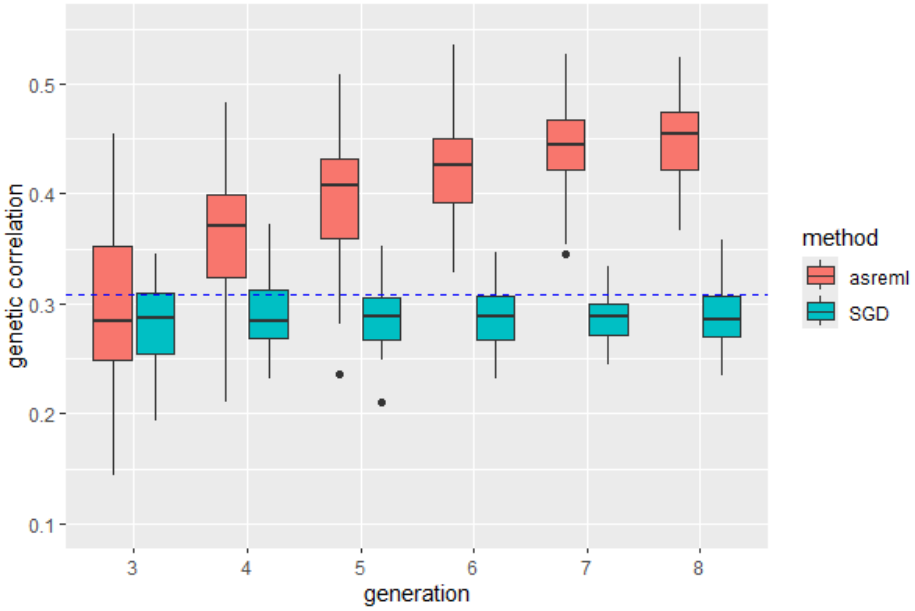


FIGURE 2 – Correlation génétique estimée en utilisant l'approche REML ou en utilisant l'approche SGD, de la génération G_3 la génération G_8

Meyer, K. (1985), 'Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices', *Biometrics* pp. 153–165.

Mrode, R. A. (2014), *Linear models for the prediction of animal breeding values*, 3rd ed. Wallingford : CABI.

Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.

Robert, C. P., Casella, G. & Casella, G. (1999), *Monte Carlo statistical methods*, Vol. 2, Springer.

Rohmer, T., Ricard, A. & David, I. (2022), 'Copula miss-specification in REML multivariate genetic animal model estimation', *Genetics Selection Evolution* **54**(1), 36.

Sklar, M. (1959), 'Fonctions de repartition an dimensions et leurs marges', *Publ. inst. statist. univ. Paris* **8**, 229–231.

Enbis & Groupe Fiabilité et incertitudes 1

Des architectures de réseaux de neurones innovantes pour modéliser la consommation d'électricité

Yvenn Amara-Ouali [1,2]

[1] *University Paris-Saclay, LMO, France, yvenn.amara-ouali@universite-paris-saclay.fr*
[2] *EDF R&D, OSIRIS, Palaiseau, France*

Résumé. Cet article explore l'application d'architectures innovantes d'apprentissage profond pour améliorer la prévision de la demande d'électricité dans les systèmes de réseaux décentralisés, en se concentrant sur le marché français de l'énergie. Les modèles additifs généralisés (GAM), qui sont des méthodes de pointe pour la prévision de la charge électrique, ont du mal à gérer les dépendances spatiales et les interactions à haute dimension inhérentes aux réseaux modernes. Pour combler ces lacunes, nous proposons des cadres d'apprentissage profond complémentaires : Les réseaux neuronaux graphiques (GNN) pour modéliser les hiérarchies spatiales entre les régions, les réseaux neuronaux convolutionnels (CNN) multirésolution pour intégrer les données temporelles hétérogènes, et les techniques de méta-apprentissage comme le package DRAGON pour optimiser automatiquement les architectures neuronales. Nous présentons une étude de cas sur la prévision de la demande électrique nationale française de l'année 2022 à trois niveaux hiérarchiques - national, régional (12 régions) et urbain (12 villes) - en utilisant une fonction de perte simple (RMSE) et des ensembles de données ouvertes de RTE et Météo France.

Mots-clés. Prévision de la demande d'électricité, réseaux neuronaux graphiques, modèles additifs généralisés, modèles hybrides, méta-apprentissage.

Abstract. This paper explores the application of innovative deep learning architectures to enhance electricity demand forecasting in decentralized grid systems, with a focus on the French energy market. Generalised Additive Models (GAMs), which are state of the art methods for electricity load forecasting, struggle with spatial dependencies and high-dimensional interactions inherent in modern grids. To address these gaps, we propose complementary deep learning frameworks: Graph Neural Networks (GNNs) for modeling spatial hierarchies across regions, multi-resolution Convolutional Neural Networks (CNNs) for integrating heterogeneous temporal data, and meta-learning techniques like the DRAGON framework to optimize neural architectures automatically. A case study forecasts 2022 French national electricity load at three hierarchical levels—national, regional (12 regions), and city (12 cities)—using a composite loss function (RMSE) and open datasets from RTE and Météo France.

Keywords. Electricity demand forecasting, Graph Neural Networks, Generalized Additive Models, hybrid models, meta-learning.

1 Introduction

Electricity markets are undergoing a transformative shift driven by the increasing integration of renewable energy sources, evolving consumer behaviors, and the growing complexity of grid operations [1]. Accurate electricity demand forecasting is critical for ensuring grid stability, optimizing energy dispatch, and enabling efficient market operations [2]. Traditional forecasting models, such as autoregressive integrated moving average (ARIMA) and linear regression, often struggle to capture the nonlinearities, temporal dependencies, and external factors influencing electricity demand, particularly in the context of rapidly changing energy landscapes [3, 4]. Generalized Additive Models (GAMs) have been widely used to address some of these limitations by modeling nonlinear relationships through smooth functions of predictors, offering interpretability and flexibility [5, 6]. However, GAMs may still fall short in capturing complex interactions and high-dimensional dependencies inherent in modern electricity markets. Novel deep learning architectures, with their ability to model complex patterns and adapt to dynamic environments, offer a promising avenue to address these challenges. For instance, attention mechanisms and transformer-based models have demonstrated superior performance in capturing long-term dependencies in time series data [7], while graph neural networks (GNNs) have shown potential in modeling spatial dependencies in energy systems [8]. By leveraging these advancements, researchers can develop more robust and accurate forecasting tools, ultimately supporting the transition to a sustainable and resilient energy future [9].

2 Background

In this section, we provide a short background on some state of the art methods for modelling electricity demand.

Generalized Additive Models Generalized Additive Models (GAMs) are a flexible class of statistical models that extend Generalized Linear Models (GLMs) by allowing for nonlinear relationships between predictors and the response variable. GAMs achieve this by modeling the response variable as a sum of smooth functions of the predictors, rather than assuming a strictly linear relationship. Mathematically, a GAM can be expressed as:

$$g(\mathbb{E}[Y]) = \beta_0 + \sum_{j=1}^{K} f_j(x_j) + \epsilon \tag{1}$$

where:

- Y is the response variable (e.g., electricity demand),
- $g(\cdot)$ is the link function (e.g., identity, log, or logistic),
- β_0 is the intercept term,

- $f_j(x_j)$ are smooth functions (e.g., splines) applied to the predictors x_j,
- $x_1, x_2, \ldots, x_K$ are the predictor variables (e.g., temperature, time of day, day of week).

The smooth functions $f_i(x_i)$ are typically represented using basis functions, such as cubic splines or thin-plate splines, which allow for flexible modeling of nonlinear effects. The choice of basis functions and the degree of smoothness are controlled by regularization parameters, which can be optimized using techniques like cross-validation or maximum likelihood estimation [5, 6]. GAMs are particularly well-suited for electricity demand forecasting due to their ability to capture nonlinear relationships between demand and exogenous factors such as temperature, humidity, and calendar effects. For example, the relationship between electricity demand and temperature is often nonlinear, with demand peaking during extreme hot or cold conditions due to heating or cooling needs. GAMs can model these effects explicitly, providing interpretable insights into the drivers of electricity demand [10].

Application of GAMs to Electricity Demand Forecasting In the context of electricity demand forecasting, GAMs have been widely used to model the complex interplay between demand and external factors. For instance, [10] demonstrated the effectiveness of GAMs in capturing the nonlinear effects of temperature and seasonal variations on electricity demand. Similarly, [11] applied GAMs to model daily and weekly patterns in electricity consumption, highlighting their ability to handle multiple temporal scales simultaneously. One of the key advantages of GAMs is their interpretability. Unlike "black-box" models such as deep neural networks, GAMs provide a transparent representation of the relationship between predictors and the response variable. This makes them particularly useful for decision-making in energy systems, where understanding the drivers of demand is critical for planning and operations [6].

What about deep learning? Deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown remarkable success in short-term load forecasting due to their ability to capture complex temporal dependencies and nonlinear patterns in high-resolution data. These models excel at automatically extracting features from raw input data, making them particularly effective for tasks like electricity demand forecasting, where the relationship between predictors and the target variable can be highly intricate. However, while deep learning models often perform similarly to Generalized Additive Models (GAMs) in terms of predictive accuracy, they do not always outperform them. GAMs, with their interpretability and ability to model nonlinear relationships explicitly, remain a strong baseline, and show great flexibility even in multi-resolution settings [12]. Nevertheless, the complementary strengths of deep learning models—such as their capacity to learn hierarchical representations and handle high-dimensional interactions—suggest that novel architectures could offer alternative ways to represent the forecasting problem differently than GAMs. In expert aggregation frameworks [13], combining models that have learned different aspects of the data (e.g., GAMs for interpretability and deep learning for capturing complex dependencies) can lead to more robust and accurate forecasts. This high-

lights the need for continued innovation in deep learning architectures to explore new representations and improve the diversity of expert models in ensemble forecasting systems.

3 Relevant deep learning architectures

Deep learning and graphs [14] proposes a deep learning framework using Graph Neural Networks (GNNs) to model spatial dependencies in electricity demand, contrasting with traditional Generalized Additive Models (GAMs). One limation of GAMs is that they struggle to capture spatial dependencies between the different nodes of a decentralized grid. GNNs address this limitation by explicitly modeling interactions between nodes (e.g., regions) via graph structures, leveraging message-passing mechanisms to aggregate neighborhood information [15]. While GAMs remain superior in standalone accuracy (e.g., 1.48% MAPE vs. 1.78% for GNNs in real data), expert aggregation combining GAMs and GNNs achieved the best results (1.13% MAPE), demonstrating that deep learning complements—rather than replaces—GAMs by addressing their spatial modeling limitations. Deep learning architectures like GNNs provide a novel paradigm for electricity forecasting by learning latent spatial structures and dynamic interactions, offering complementary strengths to GAMs' interpretable additive framework. Hybrid approaches that aggregate these"experts" are key to balancing accuracy, interpretability, and adaptability in modern energy systems.

Optimal Deep Learning Architectures Neural Architecture Search (NAS) and Hyperparameter Optimization (HPO) have become crucial methodologies for improving the efficiency and predictive power of deep neural networks (DNNs) in time series forecasting. The DRAGON framework [16] introduces a novel evolutionary approach for optimizing DNN architectures by encoding them as Directed Acyclic Graphs (DAGs), allowing flexible combinations of convolutional, recurrent, and attention-based layers. This methodology has been shown to outperform both handcrafted models and existing AutoML techniques on various forecasting tasks.

In the domain of electricity load forecasting, where accurate predictions are essential for grid stability and energy management, DRAGON offers a compelling alternative to traditional methods such as GAMs. The adaptability of DRAGON enables automatic discovery of architectures tailored to specific datasets, resulting in improved forecasting accuracy and robustness. Moreover, by leveraging an evolutionary search process, DRAGON explores a broader function space than manually designed models, addressing the limitations of static model selection and enabling the discovery of novel deep learning structures suited to the dynamic nature of electricity consumption patterns.

4 Case Study - Forecasting French National Load at different levels of Hierarchy

In this section we present a case study for deep learning architectures applied to electricity load forecasting in a decentralised grid paradigm.

4.1 Problem Formulation

Target vector The goal is to forecast electricity consumption for the year 2022 at different levels of the decentralised grid. In particular, the target vector $\boldsymbol{y} = (y_k)_{k\in\{1...25\}}$ contains 25 coordinates, 1 for France, 12 for the 12 administrative regions of metropolitan France (see Figure 1) and 12 for 12 main cities (see Figure 2).

Loss function The loss function is the sum of the root-mean squared errors (RMSE) of the forecasted values for each node :

$$\ell = \sum_{k=1}^{25} \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_k - \widehat{y}_k)^2} \tag{2}$$

with N the number of timesteps forecasted and $\widehat{\boldsymbol{y}} = (\widehat{y}_k)_{k\in\{1...25\}}$ the forecasts for each node.

4.2 Datasets

Electricity Demand The electricity demand datasets are open and available and provided by RTE at the following link.

Weather Observed The weather demand dataset is open and available and provided by Meteo France at the following link.

Codabench Platform We made a platform available to recuperate the data and submit forecasts for the year 2022 with a leaderboard available at the following link.

4.3 Experiments and Results

The final analysis of the ongoing load forecasting experiments is still in progress. While the detailed findings are not yet available, we are pleased to confirm that the comprehensive results, including comparative evaluations of model performance and their implications for energy management systems, will be presented in full at the upcoming conference. We

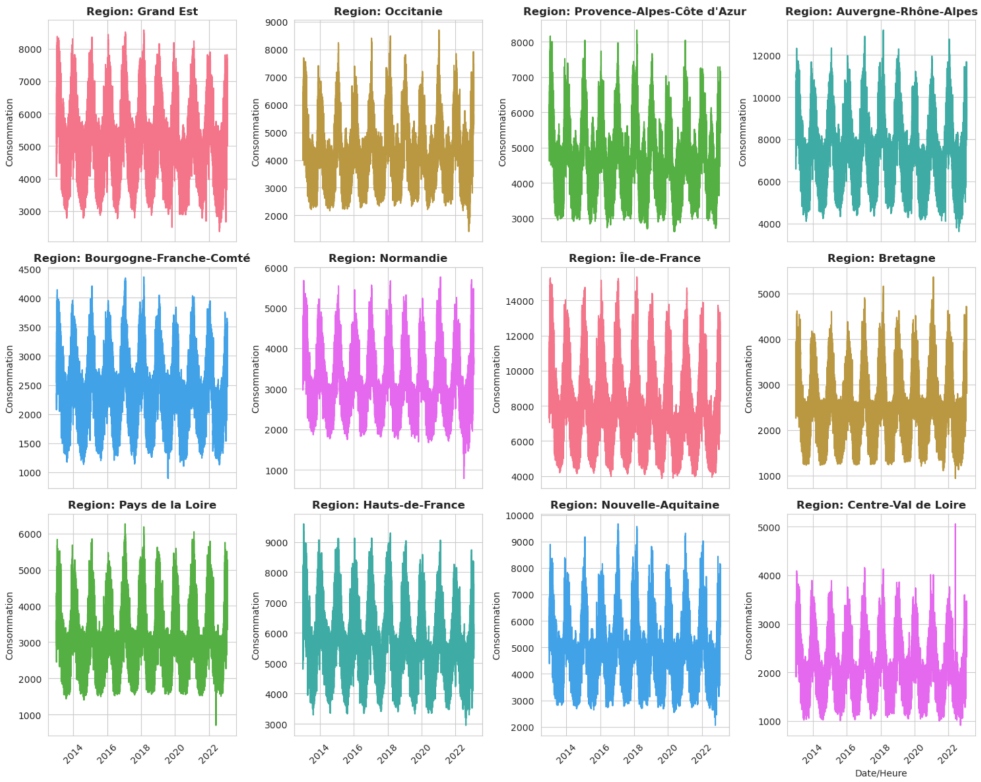


Figure 1: Load consumption in the 12 regions

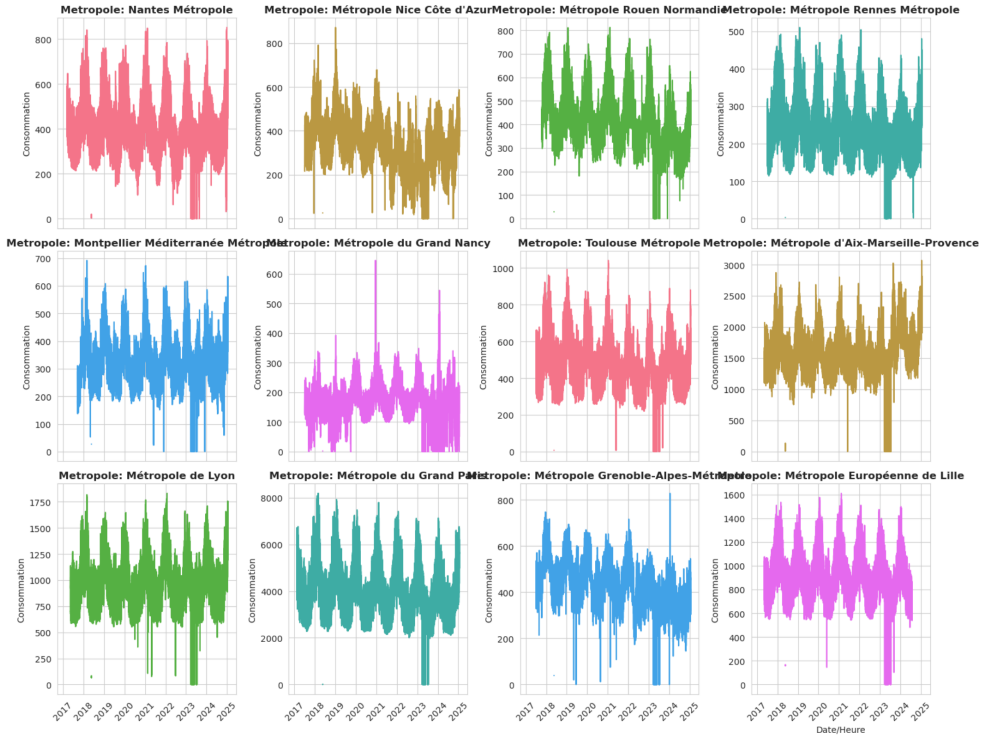


Figure 2: Load consumption in the 12 cities

appreciate your patience and look forward to sharing insights that could contribute to more efficient grid operations and sustainable energy planning.

References

[1] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, "A review of deep learning for renewable energy forecasting," *Energy Conversion and Management*, vol. 198, p. 111799, 2019.

[2] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," *International Journal of Forecasting*, vol. 30, no. 4, pp. 1030–1081, 2014.

[3] S. Hochreiter, "Long short-term memory," *Neural Computation MIT-Press*, 1997.

[4] S. Smyl, "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting," *International journal of forecasting*, vol. 36, no. 1, pp. 75–85, 2020.

[5] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*. Chapman and Hall/CRC, 1990.

[6] S. N. Wood, "Generalized additive models: An introduction with r," *Chapman and Hall/CRC*, 2017.

[7] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[8] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, *et al.*, "Relational inductive biases, deep learning, and graph networks. arxiv 2018," *arXiv preprint arXiv:1806.01261*, 2018.

[9] J. Lago, F. De Ridder, and B. De Schutter, "Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms," *Applied Energy*, vol. 221, pp. 386–405, 2018.

[10] S. Fan and R. J. Hyndman, "Short-term load forecasting based on a semi-parametric additive model," *IEEE transactions on power systems*, vol. 27, no. 1, pp. 134–141, 2011.

[11] A. M. De Livera, R. J. Hyndman, and R. D. Snyder, "Forecasting time series with complex seasonal patterns using exponential smoothing," *Journal of the American statistical association*, vol. 106, no. 496, pp. 1513–1527, 2011.

[12] Y. Amara-Ouali, M. Fasiolo, Y. Goude, and H. Yan, "Daily peak electrical load forecasting with a multi-resolution approach," *International Journal of Forecasting*, vol. 39, no. 4, pp. 1272–1286, 2023.

[13] P. Gaillard, Y. Goude, and R. Nedellec, "Additive models and robust aggregation for gefcom2014 probabilistic electric load and electricity price forecasting," *International Journal of forecasting*, vol. 32, no. 3, pp. 1038–1050, 2016.

[14] E. Campagne, Y. Amara-Ouali, Y. Goude, and A. Kalogeratos, "Leveraging graph neural networks to forecast electricity consumption," *arXiv preprint arXiv:2408.17366*, 2024.

[15] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[16] J. Keisler, E.-G. Talbi, S. Claudel, and G. Cabriel, "An algorithmic framework for the optimization of deep neural networks architectures and hyperparameters," *Journal of Machine Learning Research*, vol. 25, pp. 1–33, 2024.

Etude d'indicateurs de pertinence d'une base de données expérimentales pour la validation de codes de calculs. Application sur la simulation de transitoires thermohydrauliques accidentels.

Jean Baccou [1], Tony Glantz [1], Bertrand Iooss [2] & Roman Sueur [2]

[1] *ASNR, France, jean.baccou@asnr.fr, tony.glantz@asnr.fr*
[2] *EDF R&D, France, bertrand.iooss@edf.fr, roman.sueur@edf.fr*

Résumé. Dans certaines études de sûreté nucléaire, des transitoires accidentels sont simulés par des codes numériques. Ceux-ci comportent de nombreux paramètres physiques souvent caractérisés par des bases d'essais expérimentaux. Ce travail a pour objectif de progresser sur le développement et l'application d'une méthode générique et formelle d'analyse de la pertinence de ces bases. Les propriétés de représentativité et d'exhaustivité sont explicités, ainsi que la démarche méthodologique pour leur analyse. Celle-ci s'appuie sur l'utilisation de méthodes multicritères pour la représentativité et sur la construction d'un nouvel indicateur combinant métrique "space-filling" et critère statistique pour l'exhaustivité. Une application en sûreté nucléaire est ensuite fournie et montre notamment que l'indicateur d'exhaustivité permet d'évaluer si une base expérimentale est exhaustive. Pour améliorer l'exhaustivité, l'enrichissement d'une base existante par des nouveaux tests est également illustré.

Mots-clés. Analyse multicritères, Bases expérimentales, Représentativité, Exhaustivité, Métrique "Space-filling"

Abstract. In some safety studies, accidental transients are simulated by computer codes, which contain many physical parameters characterized by experimental data. The objective of this work is to progress on the development and application of a generic and formal approach to analyze the adequacy of a database. The two main properties (representativeness and completeness) that characterize the adequacy are first recalled as well as the methodological tools to perform their analyses. For representativeness, it relies on multicriteria methods. For completness, it exploits a new indicator that combines space-filling metrics and statistical criteria. An application on real data in nuclear safety analysis is then provided and shows that the completeness indicator can be used to evaluate if a given database is complete. Finally, database augmentation is applied to improve the completeness of a database.

Keywords. Multicriteria analysis, Experimental bases, Representativeness, Completeness, Space-filling metrics,

1 Introduction

L'étude de la sûreté des réacteurs du parc de production d'EDF implique l'examen de différents scénarios d'accidents afin de justifier du respect des critères réglementaires auprès de l'autorité en charge du contrôle des activités nucléaires. Les transitoires accidentels sont

simulés à l'aide de codes numériques modélisant les phénomènes physiques mis en jeu dans les scénarios étudiés. En particulier, les codes de thermo-hydraulique qui permettent de modéliser et simuler les écoulements dans les différents circuits et composants de la chaudière des réacteurs. Ils comportent néanmoins de nombreux paramètres physiques intervenant dans les équations modélisant les écoulements. Ces paramètres et leurs incertitudes (caractérisées par des lois de probabilité) sont déterminés par des essais réalisés sur des dispositifs expérimentaux reproduisant, dans une certaine mesure, les configurations d'écoulement rencontrées sur un réacteur réel. Ces essais étant, par ailleurs, coûteux et complexes à mettre en œuvre, un même dispositif expérimental est généralement utilisé pour l'estimation d'un nombre important de paramètres physiques intervenant dans le code numérique. Se pose alors la question de l'adéquation entre les essais réalisés et les paramètres dont on cherche à caractériser les incertitudes (Baccou et al., 2020).

Les incertitudes pouvant impacter significativement les résultats de simulation, on cherche à évaluer de la manière la plus objective et rigoureuse possible la pertinence d'une base de données expérimentale à une modélisation particulière. Baccou et al. (2018) proposent d'évaluer cette pertinence par un indicateur de représentativité et un indicateur d'exhaustivité de la base. De premières constructions et applications de la représentativité, exploitant une méthode d'aide à la décision multicritères (Roy, 1996), ont été réalisées dans le cadre du projet international ATRIUM (Baccou et al., 2024). Ce travail[1] en présente la suite.

Les deux sections suivantes explicitent les notions de représentativité (ainsi que la démarche suivie pour analyser la représentativité par approches multicritères) et d'exhaustivité (avec un nouvel indicateur d'exhaustivité). Les deux dernières sections décrivent, respectivement, une application de ces développements dans le cadre de la sûreté nucléaire et une conclusion.

2 Analyse de la représentativité

La représentativité d'une base de données désigne sa capacité à fournir des informations exploitables pour la modélisation et la simulation des scénarios étudiés. Par exemple, si une base est issue d'une installation expérimentale de géométrie similaire à celle du système industriel visé, on parlera de représentativité géométrique. La représentativité est cependant un problème multifactoriel. Pour quantifier un indicateur de représentativité, on s'oriente alors vers l'analyse décisionnelle multicritères (Roy, 1996). Ce type d'approches nécessite d'abord l'introduction d'une famille de critères non redondants et traduisant tous les aspects de la représentativité d'une base. Puis, la représentativité de chaque expérience est notée par critère, chaque critère ayant un poids d'importance. Cela conduit à une table de décision construite en relation étroite avec les experts. Ensuite, une règle de décision est appliquée à la table en combinant les évaluations par critère et les pondérations pour affecter à chaque expérience une note de représentativité globale ou pour classer les expériences des plus représentatives aux moins représentatives.

La plupart des méthodes multicritères sont basées sur des relations binaires (relation du

[1]Réalisé dans le cadre du projet QUTHY du programme NEEDS (accord de coopération géré par le CNRS dans le domaine du nucléaire entre CNRS, Andra, BRGM, CEA, EDF, Framatome, IRSN et Orano).

type “est plus représentative que”) qui permettent de comparer les bases expérimentales entre elles. Il en existe de nombreuses dans la littérature (Saaty, 1990; Roy, 1996) conduisant à un indicateur unique de synthèse ou directement à un classement (pouvant être partiel) des bases selon leur représentativité. Dans ce travail, on utilise la méthode multicritère AHP (“Analytic Hierarchy Process”) de Saaty (1990). Il est demandé à l'expert de considérer chaque paire de critères et chaque paire de bases puis de fournir leur ratio d'importance ou de représentativité selon une échelle numérique allant de 1 à 9. Un calcul de vecteurs propres permet ensuite d'obtenir le vecteur d'importance des critères et le vecteur de représentativité des bases pour chaque critère. Enfin, un indicateur unique de synthèse de la représentativité est calculé par moyenne pondérée des représentativités par critère.

Cette approche, basée sur un cadre mathématique, est aisée à mettre en œuvre en pratique puisqu'elle requiert de simples calculs matriciels. Toutefois, le cadre formel de la méthode implique une hypothèse forte sur la structure de la matrice construite à partir des comparaisons par paires. La méthode demande aussi à l'expert de se prononcer sur toutes les comparaisons par paires ce qui peut être contraignant (e.g. dans le cas de bases incomparables).

3 Analyse de l'exhaustivité

L'exhaustivité d'une base de données concerne la répartition spatiale de ses points dans le domaine physique, en premier lieu la capacité de la base à couvrir les plages de variation des quantités physiques dont dépendent les modèles. Obtenues par avis d'expert, ces plages définissent le domaine physique de validation. L'évaluation de l'exhaustivité consiste donc à estimer un degré de couverture de ce domaine par les points expérimentaux tout en favorisant les ensembles de points permettant d'explorer tout le domaine de validation sans privilégier certaines régions. Des métriques de type “space filling”, utilisées classiquement dans le domaine de l'expérimentation numérique (Da Veiga et al., 2021), sont alors pertinentes pour construire un indicateur d'exhaustivité. Toutefois, la plupart d'entre elles ne permettent pas d'évaluer si une répartition de points est suffisamment exhaustive et sont limitées à des domaines de type hypercubes. Cela ne correspond pas toujours à la forme des domaines des bases expérimentales, certaines régions de l'hypercube pouvant être exclues car non physiques.

Afin de contourner ces difficultés, nous proposons la construction d'un indicateur en plusieurs étapes. Cela nécessite d'avoir à sa disposition un ensemble de N bases fictives de référence jugées exhaustives par l'expert. Si $\{B_i^{\text{ref}}\}_{i=1,\dots,N}$ désigne cet ensemble de bases, la construction de l'indicateur d'exhaustivité $\mathcal{C}^{B^{\text{ref}}}$ repose sur les deux ingrédients suivants :

1) Choix d'une métrique “space-filling” notée M. Nous en proposons deux différentes : a) IMSE : moyenne sur tout le domaine de validation de l'erreur quadratique moyenne issue de l'ajustement d'un modèle de krigeage (Sacks et al., 1989) ; b) MST : construction de l'arbre de longueur minimale (cf, e.g., Da Veiga et al., 2021) joignant les points expérimentaux. Dans ce cas, la métrique est le rapport entre l'écart type et la moyenne de la longueur des arêtes de l'arbre.

2) Comparaison entre $M(B^\star)$ et $\{M(B_i^{\text{ref}})\}_{i=1,\dots,N}$ ($B^\star$ base à analyser) via un critère (Liu et al., 2011). Nous explorons deux types de critères : a) $\mathcal{C}_0^{B^{\text{ref}}}$ qui se base sur la vraisemblance

de $M(B^\star)$ vis-à-vis de la loi construite à partir des valeurs de référence de la métrique $\{M(B_i^{\text{ref}})\}_{i=1,\dots,N}$ (Marrel et al., 2024). b) $\mathcal{C}_1^{B^{\text{ref}}}$ qui mesure la centralité de $M(B^\star)$ parmi l'échantillon de valeurs de référence $\left(M(B_1^{\text{ref}}),\dots,M(B_N^{\text{ref}})\right)$. Si M^{ref} est la variable aléatoire associée à la loi construite à partir de $\{M(B_i^{\text{ref}})\}_{i=1,\dots,N}$, les deux indicateurs s'écrivent

$$\mathcal{C}_0^{B^{\text{ref}}}(B^\star) = \text{Proba}[f_{M^{\text{ref}}}\left(M^{\text{ref}}\right) \leq f_{M^{\text{ref}}}\left(M(B^\star)\right)], \quad (1)$$

$$\mathcal{C}_1^{B^{\text{ref}}}(B^\star) = 1 - 2|F_{M^{\text{ref}}}\left(M(B^\star)\right) - 0.5|. \quad (2)$$

où $f_{M^{\text{ref}}}$ et $F_{M^{\text{ref}}}$ sont la densité et la fonction de répartition de M^{ref}. Notons qu'une valeur proche de un de ces indicateurs traduit une base exhaustive.

La figure 1 et la table 1 donnent deux exemples d'application de ces indicateurs pour un domaine de validation triangulaire. Les bases de référence sont construites en appliquant la méthode des K-means (avec différents initiateurs) à un ensemble de 5000 points répartis uniformément dans le triangle. Une covariance de type Matérn-3/2 est ajustée à partir de la mesure expérimentale obtenue sur ce cas jouet par la fonction de Branin. Dans le cas de $B_0^\star$ qui ne présente pas de cluster de points, les deux indicateurs donnent des résultats proches quelle que soit la métrique choisie. Par contre, l'analyse de $B_1^\star$ montre un désaccord entre les indicateurs basés sur le MST et sur l'IMSE. En effet, la présence du cluster fait chuter la valeur de l'indicateur intégrant la métrique MST (la base n'est plus exhaustive) alors que l'indicateur basé sur l'IMSE est moins impacté.

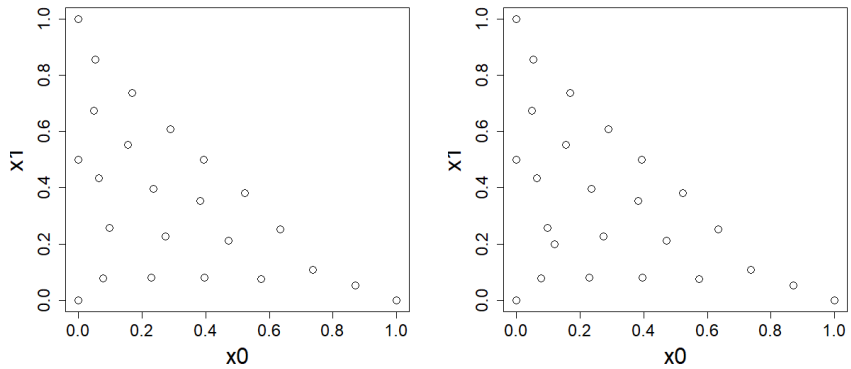


Figure 1: Exemples de deux bases de données (notées $B_0^\star$ et $B_1^\star$) à analyser.

Métrique	Base $B_0^\star$	Base $B_1^\star$
MST	(0.56,0.77)	(0.01,0.02)
IMSE	(0.94,0.82)	(0.53,0.61)

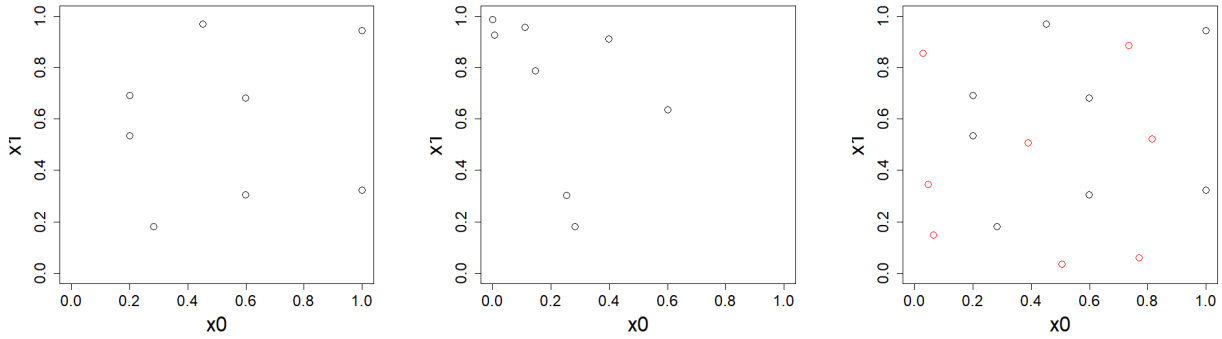
Table 1: Valeurs des critères d'exhaustivité pour les bases $B_0^\star$ et $B_1^\star$. La première valeur est associée à $\mathcal{C}_0^{B^{\text{ref}}}$, la seconde à $\mathcal{C}_1^{B^{\text{ref}}}$.

4 Application en sûreté nucléaire

Cette application est issue du projet international ATRIUM (Ghione et al., 2023; Baccou et al., 2024) qui a pour objectif de quantifier les incertitudes dans le cadre de la simulation

numérique de phénomènes physiques d'intérêt ayant lieu pendant un accident de perte de réfrigérant primaire par une brèche de taille intermédiaire dans un réacteur nucléaire. Huit bases expérimentales sont à analyser, elles sont notées S_2, S_3, S_4, SMD_{div}, SMD_{exp}, M_{13}, M_{17}, M_{24}. Les expériences de ces bases sont décrites par des variables (explicatives, ici notées x0 et x1) déterminant les conditions physiques de réalisation de chaque essai (variables d'entrée du modèle physique à valider) ainsi que par une quantité physique d'intérêt (sortie du modèle physique, ici un débit critique à la brèche). L'analyse consiste d'abord à identifier les deix bases expérimentales les plus représentatives (par rapport à une expérience cible proche du cas réacteur) puis d'étudier l'exhaustivité de plusieurs ensembles de points expérimentaux provenant de la combinaison de ces deux bases. Dans Baccou et al. (2024), trois critères de représentativité ont été proposés. Le premier concerne la représentativité de chaque base expérimentale par rapport à l'expérience cible en terme de géométrie et de conditions thermohydrauliques. Le deuxième se focalise sur la qualité des données expérimentales. Le troisième porte sur la capacité de chaque base à couvrir les phénomènes physiques simulés par le code et à les traiter séparément. Il a alors été montré que SMD_{div}, SMD_{exp} sont les deux bases les plus représentatives.

En combinant les tests expérimentaux de ces bases représentatives, nous évaluons dans la figure 2 ((a) et (b)) l'exhaustivité de deux ensembles de tests expérimentaux issus de cette combinaison. Le domaine de validation est ici carré. De plus, les bases de référence sont obtenues en générant 1000 plans LHS optimisés. Enfin, la métrique est l'IMSE (avec une fonction de covariance isotrope Matérn 3/2 ajustée à partir des mesures expérimentales du débit critique) et les critères d'exhaustivité sont donnés par les expressions (1) et (2).



(a)
$\mathcal{C}_0^{B^{\text{ref}}}(B^\star) = 7.5e-3$
$\mathcal{C}_1^{B^{\text{ref}}}(B^\star) = 8e-3$

(b)
$\mathcal{C}_1^{B^{\text{ref}}}(B^\star) = 0$
$\mathcal{C}_1^{B^{\text{ref}}}(B^\star) = 0$

(c)
$\mathcal{C}_1^{B^{\text{ref}}}(B^\star) = 0.95$

Figure 2: Exhaustivités pour deux ensembles de tests issus des bases les plus représentatives.

Les deux critères d'exhaustivité montrent que le premier ensemble de points est plus exhaustif que le second. Toutefois, la valeur de chaque critère est faible indiquant un manque d'exhaustivité. Afin de l'améliorer, une stratégie séquentielle d'enrichissement (par de nouveaux points) peut être construite. Cela fournit des nouveaux tests expérimentaux à réaliser, illustrés par les points rouges en Fig. 2 (c). Dans le cas de l'IMSE, la stratégie d'enrichissement s'appuie sur la recherche des points qui minimisent cette quantité.

5 Conclusion

Dans ce travail, la pertinence (représentativité et exhaustivité) d'une base expérimentale est étudiée en amont de l'application de méthodes inverses de quantification d'incertitudes en sûreté nucléaire. Il s'agit plus précisément de progresser sur le développement et l'application d'une méthode générique (indépendante d'un jeu de bases expérimentales particulier) et formelle (mathématisable, voire automatisable) afin de mettre à la disposition des ingénieurs des outils d'analyse. Le traitement de la représentativité s'appuie sur une approche multi-critères. L'exhaustivité est étudiée en introduisant un indicateur combinant métrique "space-filling" et critère statistique. L'utilisation de ces développements facilitera le partage des connaissances et la transparence des différences d'appréciation entre les experts.

Bibliographie

Baccou, J., Chojnacki, E., Georgenthum, V. and Marais, S. (2018), On the methodological treatment of input uncertainty quantification: illustration in the RIA framework, *ANS Best Estimate Plus Uncertainty International Conference (BEPU 2018)*, Lucca, Italy, May 13-19.

Baccou, J. et al. (2020), SAPIUM: a generic framework for a practical and transparent quantification of thermal hydraulic code model input uncertainty, *Nuclear Science and Engineering*, 194 (8-9), 721-736.

Baccou, J. et al. (2024), A systematic approach for the adequacy analysis of a set of experimental databases: Application in the framework of the ATRIUM project, *Nuclear Engineering and Design*, 421:113035.

Da Veiga, S., Gamboa, F., Iooss, B. and Prieur, C. (2021), Basics and trends in sensitivity analysis - Theory and practice in R, SIAM.

Ghione, A., Sargentini, L., Damblin, G., Fillion, P., Baccou, J., Sueur, R., Iooss, B., Petruzzi, A., Zeng, K., Zhang, J., Mendizabal, R., Skorek, T., Wu, X., Freixa, J. and Adorni M. (2023), Application of SAPIUM guidelines to Input Uncertainty Quantification: the ATRIUM project, *NURETH-20 conference*, Washington DC, USA.

Liu, Y., Chen, W., Arendt, P. and Huang, H-Z. (2011), Toward a better understanding of model validation metrics, *Journal of Mechanical Design*, 133:071005.

Marrel, A., Velardo, H. and Bouloré, A. (2024), Likelihood and depth-based criteria for comparing simulation results with experimental data, in support of validation of numerical simulators, *International Journal for Uncertainty Quantification*, 14(2):45-68.

Roy, B. (1996), *Multicriteria methodology for decision aiding*, Kluwer Academic Publishers.

Saaty, T.L. (1990), How to make a decision: the Analytical Hierarchical Process, *European Journal of Operational Research*, 48:9-26.

Sacks, J., Welch, W.J., Mitchell, T.J. and Wynn, H. (1989), Design and analysis of computer experiments, *Statistical Science*, 4(4), 409-423.

Prévision de séries temporelles sous contraintes linéaires

Nathan Doumèche [1,2] & Francis Bach [3] & Eloi Bedek [2] & Gérard Biau [1] & Claire Boyer [1] & Yannig Goude [2]

[1] *LSPM-Sorbonne Université, CNRS, France,*
{nathan.doumeche, gerard.biau, claire.boyer}@sorbonne-universite.fr
[2] *Électricité de France R&D, France,*
{nathan.doumeche, eloi.bedek, yannig.goude}@edf.fr
[3] *INRIA Paris, France,*
francis.bach@inria.fr

Résumé. La prévision de séries temporelles présente des particularités (échantillonnage irrégulier et corrélé, petits jeux de données...) limitant l'efficacité des algorithmes traditionnels d'apprentissage automatique. Diverses approches, telles les modèles additifs généralisés ou la prévision hiérarchique, parviennent à améliorer la performance des modèles en intégrant des contraintes dans les algorithmes d'apprentissage. Dans cette présentation, nous proposons un cadre unifié pour l'intégration et la combinaison de contraintes linéaires spécifiques à la prévision des séries temporelles. Dans ce cadre, nous montrons que le minimiseur exact du risque empirique contraint peut être calculé efficacement en utilisant uniquement de l'algèbre linéaire. Cette approche permet des implémentations efficaces sur carte graphique (GPU). Nous validons la méthodologie proposée par des expériences sur données réelles en prévision de demande d'électricité et de tourisme. Les algorithmes contraints sont non seulement plus interprétables que l'état-de-l'art, mais ils obtiennent aussi des performances significativement meilleures.

Mots-clés. Apprentissage statistiques informé de la physique, contraintes, prévision hiérarchique, transfert d'information, prévision de consommation électrique, prévision de flux de tourisme.

Abstract. Time series forecasting presents unique challenges that limit the effectiveness of traditional machine learning algorithms. To address these limitations, various approaches have incorporated linear constraints into learning algorithms, such as generalized additive models and hierarchical forecasting. In this paper, we propose a unified framework for integrating and combining linear constraints in time series forecasting. Within this framework, we show that the exact minimizer of the constrained empirical risk can be computed efficiently using linear algebra alone. This approach allows for highly scalable implementations optimized for GPUs. We validate the proposed methodology through extensive benchmarking on real-world tasks, including electricity demand forecasting and tourism forecasting, achieving state-of-the-art performance.

Keywords. Physics-informed machine learning, constraints, hierarchical forecasting, transfer learning, load forecasting, tourism forecasting.

1 Introduction

Apprentissage sous contraintes différentielles. L'incorporation de contraintes dans les algorithmes d'apprentissage automatique a vu un regain d'intérêt avec l'émergence du *Physics-informed machine learning* [Karniadakis et al., 2021], en particulier par l'introduction des réseaux de neurones informés par la physiques (PINNs) par Raissi et al. [2019]. Le principe des PINNs est d'incorporer un a priori physique, prenant la forme d'un équation aux dérivées partielles (EDPs), dans la fonction de coût d'un réseau de neurone. Dans un précédent article, nous avons démontré que les PINNs peuvent être victimes de surapprentissage, mais qu'au prix d'une régularisation ridge, ils parviennent à converger [Doumèche et al., 2025]. Cependant, la complexité informatique nécessaire à la convergence des PINNs est importante, et nous avons proposé une alternative efficace pour les EDPs linéaires consistant à reformuler le problème comme un problème à noyau [Doumèche et al., 2024]. Cela a aboutit à l'implémentation sur carte graphique (GPU) d'un algorithme d'optimisation efficace, qui de plus a de meilleurs performances que les PINNs pour la résolution d'EDPs linéaires [Doumèche et al., 2024].

Applications aux séries temporelles. Les séries temporelles font face à de nombreux défis structurels (changement dans la distribution des données...) qui limitent la taille des échantillons de données utiles et qui les rendent particulièrement difficiles prévoir. Aussi, il est courant que les praticiens appliquent des contraintes linéaires aux modèles pour guider l'apprentissage et ainsi améliorer les performances. Malheureusement, les séries temporelles issues du monde socio-économique ne satisfont que très rarement un jeu d'EDPs connu, et le formalisme du *Physics-informed machine learning* ne s'applique pas directement.

Contributions. Ici, nous proposons une méthode pour étendre l'algorithme de [Doumèche et al., 2024] à la modélisation de contraintes courantes en prévision de séries temporelles. Cela permet de combiner plusieurs contraintes et de les implémenter efficacement sur GPU par un algorithme que nous appelons *WeaKL*. Nous distinguons deux types de contraintes, que nous illustrons sur deux exemples avec données réelles. Les contraintes de forme consistent à restreindre la forme de la fonction de régression, et sont illustrées en prévision de consommation électrique. Les contraintes d'apprentissage consistent à guider l'apprentissage des paramètres optimaux, et sont illustrées en prévision de flux touristiques. Dans les deux cas, les méthodes contraintes obtiennent des performances de l'ordre de 10% meilleures à l'état-de-l'art. Les résultats présentés ici sont l'objet du papier Doumèche et al. [2025].

2 Contexte mathématique

Tout au long de cet exposé, nous supposons que nous avons à disposition n observations $(X_{t_1}, Y_{t_1}), \dots, (X_{t_n}, Y_{t_n})$ tirées sur $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ selon une loi inconnue. Les indices $t_1, \dots, t_n \in T$ correspondent aux instants où le processus stochastique inconnu $(X, Y) := (X_t, Y_t)_{t \in T}$ est échantillonné. Notons que, tout au long de l'article, les pas de temps ne sont pas

nécessairement réguliers sur l'ensemble des instants $T \subseteq \mathbb{R}$. L'objectif ici est d'estimer la fonction inconnue $f^\star : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ telle que $Y_t = f^\star(X_t) + \varepsilon_t$, où ε est un bruit. Sans perte de généralité, nous supposons que $X_t := (X_{1,t}, \ldots, X_{d_1,t}) \in [-\pi, \pi]^{d_1}$ et que $-\pi \leq t_1 \leq \cdots \leq t_{n+1} \leq \pi$. L'objectif est de construire un estimateur $\hat{f}$ pour $f^\star$.

L'exemple le plus simple est celui où la cible Y est une série temporelle stationnaire et régulièrement échantillonnée avec $t_j = j/n$, et où la variable décalée $X_j = Y_{j-1/n}$ est l'unique variable explicative. Dans ce cas particulier où $d_1 = d_2$, le modèle devient $Y_t = f^\star(Y_{t-1/n}) + \varepsilon_t$ et la régression se réduit à un modèle autorégressif.

3 Contraintes de forme

Le premier type de contraintes qui nous intéressent sont les contraintes de forme. Elles consistent à restreindre l'ensemble des fonctions admissibles pour construire l'estimateur $\hat{f}$. Nous considérons ici deux contraintes de forme très courantes en prévision de séries temporelles : les modèles additifs et l'adaptation en ligne.

Modèles additifs. La contrainte d'additivité consiste à supposer que $f^\star(x_1, \ldots, x_{d_1}) = \sum_{\ell=1}^{d_1} g_\ell^\star(x_\ell)$, où $g_\ell^\star : \mathbb{R} \to \mathbb{R}$ sont des fonctions univariées. Cette contrainte est très utilisée en science des données, notamment par les modèles additifs généralisés (GAMs). En effet, les modèles additifs sont interprétables parce que l'effet de chaque variable explicative x_ℓ est capturé par sa fonction correspondante $g_\ell^\star$. En outre, les effets univariés sont plus faciles à estimer que les effets multivariés (du fait du fléau de la dimension). Notre estimateur avec contrainte additive s'appelle le WeaKL additif.

Adaptation en ligne. L'adaptation en ligne consiste à mettre à jour les effets d'un modèle additif pour prendre en compte des changement dans la loi de (X_t, Y_t). Formellement, il s'agit d'apprendre les fonctions $h_\ell^\star$ telles que $f^\star(t, x_1, \ldots, x_{d_1}) = h_0^\star(t) + \sum_{\ell=1}^{d_1}(1 + h_\ell^\star(t)) g_\ell^\star(x_\ell)$. Nous appelons ce modèle le WeaKL online.

Prévision de consommation électrique. Nous comparons alors les performances de les WeaKL additif et en ligne sur le benchmark de prévision de consommation électrique de Doumèche et al. [2023]. Pour cette expérience, les performances de 10^5 formules de WeaKLs additifs et 10^4 WeaKLs en ligne ont pu être comparées sur un ensemble de validation en moins de 5 minutes sur un GPU standard (NVIDIA L4). Leur performances sont ensuite évaluées sur un ensemble de test disjoint. Les résultats sont présentés dans la Table 1. Les modèles contraints sont les plus performant et battent l'état-de-l'art de plus de 10%.

Table 1: Benchmark de prévision de consommation électrique court-terme en France

	Including holidays		Excluding holidays	
	RMSE (GW)	MAPE	RMSE (GW)	MAPE
Modèles statistiques				
Persistance (1 jour)	4.0±0.2	5.5±0.3	4.0±0.2	5.0±0.3
SARIMA	2.4±0.2	3.1±0.2	2.0±0.2	2.6±0.2
GAM	2.3±0.1	3.5±0.2	1.70±0.06	2.6±0.1
Assimilation de données				
Kalman statique	2.1±0.1	3.1±0.2	1.43±0.05	2.20±0.08
Kalman dynamique	1.4±0.1	1.9±0.1	1.10±0.04	1.58±0.05
Viking	1.5±0.1	1.8±0.1	0.98±0.04	1.33±0.04
Agrégation	1.4±0.1	1.8±0.1	0.96±0.04	1.36±0.04
Apprentissage automatique				
GAM boosting	2.6±0.2	3.7±0.2	2.3±0.1	3.3±0.2
Forêts aléatoires	2.5±0.2	3.5±0.2	2.1±0.1	3.0±0.1
Forêts aléatoires + bootstrap	2.2±0.2	3.0±0.2	1.9±0.1	2.6±0.1
Apprentissage contraint				
WeaKL additif	1.95±0.08	3.0 ±0.1	1.55±0.06	2.32±0.09
WeaKL en ligne	**1.14±0.09**	**1.5±0.1**	**0.87±0.04**	**1.17±0.05**

4 Contraintes d'apprentissage

Le second type de contraintes qui nous intéressent sont les contraintes d'apprentissage. Elles consistent à guider l'apprentissage de la fonction $\hat{f}$, sans nécessairement en contraindre la forme. L'ensemble des fonctions admissibles reste inchangé, mais nous introduisons un biais vers une certaine classe de fonctions dans l'apprentissage.

Transfert d'information. L'apprentissage par transfert permet d'exploiter les similitudes entre différentes tâches de prédiction lorsque $d_2 > 1$. Le cas le plus simple consiste à prédire plusieurs cibles $Y_1, \ldots, Y_{d_2}$ avec des caractéristiques similaires $X_1, \ldots, X_{d_2}$. Par exemple, supposons que nous cherchions à prévoir la demande d'électricité de d_2 villes. Ici, Y_ℓ est la demande d'électricité de la ville ℓ, tandis que X_ℓ est la température moyenne dans la ville ℓ. La fonction générale $f^\star$, qui estime $(Y_1, \ldots, Y_{d_2})$, peut être exprimée comme $f^\star(X) = f^\star(X_1, \ldots, X_{d_2}) = (f^{\star,1}(X_1), \ldots, f^{\star\, d_2}(X_{d_2}))$. Ici la contrainte est donc $f^{\star,1} \simeq \cdots \simeq f^{\star,d_2}$.

Modèle hiérarchique. La prévision hiérarchique consiste à prévoir plusieurs séries temporelles liées par des contraintes de sommation. Cette approche a été introduite pour prévoir le tourisme intérieur australien. En effet, le tourisme peut être analysé à différentes échelles géographiques. Par exemple, on peut considérer le nombre total de touristes en Australie, $Y_{A,t}$ à l'instant t, et le nombre de touristes $Y_{S_i,t}$ dans chacun des sept états aus-

traliens, $S_1, \ldots, S_7$. Bien entendu, le nombre total de touristes en Australie est la somme du nombre de touristes dans ses sept états, ce qui conduit à la contrainte de sommation $Y_{A,t} = \sum_{i=1}^{7} Y_{S_i,t}$. De plus, comme chaque état S_i est composé de z_i zones $Z_{i,1}, \ldots, Z_{i,z_i}$, un niveau hiérarchique supplémentaire peut être introduit. Notez que le nombre de zones dépend de l'état, pour un total de 27 zones. Il en résulte un autre ensemble de contraintes de sommation $Y_{S_i,t} = Y_{Z_{i,1},t} + \ldots + Y_{Z_{i,z_i},t}$. Il est ainsi possible de créer un ensemble de contraintes de sommation, prenant la forme d'un graphe à plusieurs niveaux.

Application à la prévision de flux touristiques. Les performances de deux modèles purement hiérarchiques (*WeaKL-BU* et *WeaKL-G*) ainsi que celles d'un modèle hiérarchique avec transfert (*WeaKL-T*) sont présentées dans la Table 2. Le modèle *WeaKL-T* obtient les meilleures performances et bat les méthodes de réconciliation de l'état-de-l'art.

Table 2: Benchmark de prévision de flux touristiques en Australie

	MSE ($\times 10^6$)					
	Australie	États	Zones	Régions	Catégories	Total
Modèles BU						
BU	5.3±0.5	2.0±0.2	1.37±0.05	**1.19±0.02**	**1.17±0.03**	11.0±0.7
WeaKL-BU	**4.5±0.5**	**1.9±0.3**	**1.34±0.05**	**1.19±0.03**	**1.17±0.03**	**10.1±0.6**
Modèles indépendants						
Indep	3.6±0.6	1.8±0.2	1.42±0.05	1.23±0.03	1.17±0.03	9.2±0.7
WeaKL-G	**3.6±0.5**	**1.8±0.2**	**1.37±0.05**	**1.18±0.03**	**1.15±0.03**	**9.0±0.7**
Toutes les données						
Rec-OLS	3.5±0.5	1.8±0.2	1.35±0.05	1.18±0.02	1.17±0.03	8.9±0.7
MinT	3.6±0.4	1.7±0.1	1.29±0.05	**1.15±0.03**	1.17±0.03	8.9±0.5
WeaKL-T	**3.1±0.3**	**1.7±0.1**	**1.27±0.05**	**1.15±0.02**	**1.12±0.03**	**8.3±0.4**

Bibliographie

N. Doumèche, Y. Allioux, Y. Goude, et S. Rubrichi. Human spatial dynamics for electricity demand forecasting: the case of france during the 2022 energy crisis. *arXiv preprint arXiv:2309.16238*, 2023.

N. Doumèche, F. Bach, G. Biau, et C. Boyer. Physics-informed machine learning as a kernel method. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1399–1450. PMLR, 2024.

N. Doumèche, G. Biau, et C. Boyer. Convergence and error analysis of PINNs. *Bernoulli, in press*, 2025.

N. Doumèche, F. Bach, G. Biau, et C. Boyer. Physics-informed kernel learning. *arXiv:2409.13786*, 2024.

N. Doumèche, F. Bach, E. Bedek, G. Biau, C. Boyer, et Y. Goude. Forecasting time series with constraint. *À paraître*, 2025.

G.E. Karniadakis, I.G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, et L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021.

M. Raissi, P. Perdikaris, et G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Intégration du machine learning dans les processus de prévision de la demande: le cas de Decathlon

Vianney Taquet*[1], Raphaël Nedellec*[1], Antoine Schwartz*†[1], and Vianney Bruned*‡[1]

[1]Decathlon – - – France

Résumé

Decathlon est une entreprise française de distribution d'articles de sports, présente dans une cinquantaine de pays, et commercialisant plusieurs dizaines de milliers de références notamment issues de ses 80 marques. A ce titre, les acteurs de la supply chain de l'entreprise se doivent de prédire au plus juste la demande de nos clients afin d'optimiser les commandes des articles auprès de nos fournisseurs et la gestion des stocks dans nos entrepôts et nos magasins.
Depuis trois ans, Decathlon a entrepris de remplacer son module de prévision de demande historique par une solution "maison" à l'état de l'art se reposant sur des modèles de machine et deep learning. Le développement de la solution s'est ainsi effectuée en deux temps:

- Solution 1.0 se reposant sur une reconstruction de l'historique de ventes et sur une combinaison d'un algorithme statistique (Holt-Winters) et d'un algorithme de deep learning (DeepAR), assurant un bon compromis entre réactivité court-terme et stabilité long-terme.
- Solution 2.0 se reposant sur une agrégation online d'experts provenant de multiples algorithmes de deep learning, tels que TFT, ou NBEATSx, et statistiques, comme les modèles additifs généralisés, s'adaptant aux différentes typologies temporelles de ventes des références de Decathlon.

Après avoir introduit le contexte et l'historique du projet, nous présenterons en détails la solution technique et les choix algorithmiques avant de terminer par un résumé des résultats et de la valeur générée par la solution proposée.

Mots-Clés: prévisions temporelles, prévisions de la demande, séries temporelles, supply chain

*Intervenant
†Auteur correspondant: antoine.schwartz@decathlon.com
‡Auteur correspondant: vianney.bruned@decathlon.com

Enbis & Groupe Fiabilité et incertitudes 2

Modelling and estimation of censored semi-Markov chains

Vlad Stefan Barbu [1,2] &Florian Lecocq [1] & Nicolas Vergne [1]

[1] *Laboratoire de Mathématique Raphaël Salem, Université de Rouen Normandie, Avenue de l'Université, BP.12, 76801 Saint-Étienne du Rouvray, Rouen, France, nicolas.vergne@univ-rouen.fr* [1] *Laboratoire de Mathématique Raphaël Salem, Université de Rouen Normandie, Avenue de l'Université, BP.12, 76801 Saint-Étienne du Rouvray, Rouen, France, florian.lecocq@oulook.com*

[2] *Centre for Demographic Research "Vladimir Trebici", "Costin C. Kiritescu" National Institute of Economic Research of Romanian Academy, Romania, barbu@univ-rouen.fr*

Résumé. Ce travail traite du problème de l'estimation paramétrique des chaînes semi-markoviennes, qui représentent une généralisation importante des chaînes de Markov et des chaînes de renouvellement. Plusieurs cadres statistiques sont envisagés : censure au début et/ou à la fin des trajectoires, pas de censure, une ou plusieurs trajectoires. Nous présentons ici le cas des chaînes semi-markoviennes générales, mais aussi trois autres cas particuliers, qui peuvent être importants d'un point de vue pratique, dans certaines applications spécifiques.

Mots-clés. Chaînes semi-markoviennes, estimation paramétrique, estimation du maximum de vraisemblance, censure, modélisation statistique.

Abstract.

This work is concerned with the problem of parametric estimation of semi-Markov chains, that represent an important generalization of Markov and renewal chains. Several statistical frameworks are considered: censoring at the beginning and/or at the end of sample paths, no censoring, one or several trajectories. We present here the case of general semi-Markov chains, but also three other particular cases, that can be important from practical point of view, in some specific applications.

Keywords. semi-Markov chains, parametric estimation, maximum likelihood estimation, censoring, sta- tistical modeling

1 Preliminaries

Let us consider a random system with finite state space $E = \{1, \ldots, s\}$, $s < \infty$. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and assume that the time evolution of the system is governed by a stochastic process $Y = (Y_k)_{k\in\mathbb{N}^*}$, defined on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in E; that is to say that Y_k gives the state of the system at time k. Let $T = (T_m)_{m\in\mathbb{N}^*}$, defined on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $\mathbb{N}$, be the successive time points when state changes in $(Y_k)_{k\in\mathbb{N}^*}$ occur (the jump times) and let also $J = (J_m)_{m\in\mathbb{N}^*}$, defined on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in E, be the successively visited

states at these time points. Set also $X = (X_m)_{m\in\mathbb{N}^*}$ for the successive sojourn times in the visited states. Thus, $X_{m+1} = T_{m+1} - T_m$, $m \in \mathbb{N}^*$, and, by convention, we set $X_1 = T_1 = 0$. The relation between the process Y and the process J of the successively visited states is given by $Y_k = J_{N(k)}$, or, equivalently, $J_m = Y_{T_m}, m, k \in \mathbb{N}$, where

$$N(k) := \max\{m \in \mathbb{N} \mid T_m \leq k\} \tag{1}$$

is the discrete-time counting process of the number of jumps in $[1, k] \subset \mathbb{N}$.

Definition 1 (semi-Markov chain SMC and Markov renewal chain MRC) *If we have*

$$\begin{aligned} & \mathbb{P}(J_{m+1} = j, T_{m+1} - T_m = k | J_m = i, J_{m-1}, \ldots, J_1, T_m, \ldots, T_1) \\ = \; & \mathbb{P}(J_{m+1} = j, T_{m+1} - T_m = k | J_m = i), \end{aligned} \tag{2}$$

then $Y = (Y_k)_k$ is called a semi-Markov chain *(SMC) and $(J, T) = (J_m, T_m)_m$ is called a* Markov renewal chain *(MRC).*

All along this paper, we assume that the MRC or SMC are homogeneous with respect to the time in the sense that Equation (2) is independent of m.

Note that if (J, T) is a MRC, then it can be proved that $J = (J_m)_{m\in\mathbb{N}^*}$ is a Markov chain with state space E, called the *embedded Markov chain* of the MRC (J, T) (or of the SMC Y).

Definition 2 *For a semi-Markov chain, we define:*

- *the* semi-Markov kernel $(q_{ij}(k))_{i,j\in E,k\in\mathbb{N}}$, $q_{ij}(k) = \mathbb{P}(J_{m+1} = j, T_{m+1} - T_m = k | J_m = i)$;
- *the* initial distribution $(\mu_i)_{i\in E}$, $\mu_i = \mathbb{P}(J_1 = i) = \mathbb{P}(Y_1 = i)$;
- *the* transition matrix $(p_{ij})_{i,j\in E}$ *of the embedded Markov chain* $J = (J_m)_m$, $p_{ij} = \mathbb{P}(J_{m+1} = j | J_m = i)$;
- *the* conditional sojourn time distributions $(f_{ij}(k))_{i,j\in E,k\in\mathbb{N}}$, $f_{ij}(k) = \mathbb{P}(T_{m+1} - T_m = k | J_m = i, J_{m+1} = j)$.

Note that

$$q_{ij}(k) = p_{ij} f_{ij}(k). \tag{3}$$

Clearly, a semi-Markov chain is uniquely determined a.s. by an initial distribution $(\mu_i)_{i\in E}$ and a semi-Markov kernel $(q_{ij}(k))_{i,j\in E,k\in\mathbb{N}}$ or, equivalently, by an initial distribution $(\mu_i)_{i\in E}$, a Markov transition matrix $(p_{ij})_{i,j\in E}$ and conditional sojourn time distributions $(f_{ij}(k))_{i,j\in E,k\in\mathbb{N}}$.

Another assumptions we do are: (i) We do not allow transitions to the same state, i.e., $p_{ii} = 0$ for all $i \in E$, or equivalently $q_{ii}(k) = 0$, for all $i \in E$, $k \in \mathbb{N}$; (ii) We assume that

there are not instantaneous transitions, that is $q_{ij}(0) \equiv 0$ or equivalently $f_{ij}(0) \equiv 0$ for all $i, j \in E$; note that this implies that T is a strictly increasing sequence.

For the conditional sojourn time distributions, one can consider the associated cumulative distribution function defined by

$$F_{ij}(k) := \mathbb{P}(T_{m+1} - T_m \leq k | J_m = i, J_{m+1} = j) = \sum_{t=1}^{k} f_{ij}(t).$$

For any distribution function $F(\cdot)$ we can consider the associated survival / reliability function defined by

$$\overline{F}(k) := 1 - F(k).$$

Consequently, we have

$$\overline{F}_{ij}(k) := \mathbb{P}(T_{m+1} - T_m > k | J_m = i, J_{m+1} = j) = 1 - \sum_{t=1}^{k} f_{ij}(t) = \sum_{t=k+1}^{\infty} f_{ij}(t). \tag{4}$$

We have considered up to here general semi-Markov models with the semi-kernel of the type given in (3). Particular types of this semi-Markov model can be taken into account, by considering particular cases of holding time distributions $f_{ij}(k)$, where these distributions depend only on the current state i, or only on the next state j, or neither on i nor on j. Consequently, one can define the semi-Markov kernel by:

$$\begin{aligned}
&q_{ij}(k) := p_{ij} f_{i\bullet}(k), \text{ where} && (5)\\
&f_{i\bullet}(k) = \mathbb{P}(T_{m+1} - T_m = k | J_m = i) = \sum_{v \in E} p_{iv} f_{iv}(k),\\
&q_{ij}(k) := p_{ij} f_{\bullet j}(k), \text{ where} && (6)\\
&f_{\bullet j}(k) = \mathbb{P}(T_{m+1} - T_m = k | J_{m+1} = j),\\
&q_{ij}(k) := p_{ij} f(k), \text{ where} && (7)\\
&f(k) = \mathbb{P}(T_{m+1} - T_m = k).
\end{aligned}$$

Note that one can introduce the distribution functions associated with the particular sojourn distributions $f_{i\bullet}(k)$, $f_{\bullet j}(k)$ or $f(k)$, defined respectively by

$$F_{i\bullet}(k) := \mathbb{P}(T_{m+1} - T_m \leq k | J_m = i) = \sum_{t=1}^{k} f_{i\bullet}(t) = \sum_{t=1}^{k} \sum_{v \in E} p_{iv} f_{iv}(t),$$

$$F_{\bullet j}(k) := \mathbb{P}(T_{m+1} - T_m \leq k | J_{m+1} = j) = \sum_{t=1}^{k} f_{\bullet j}(t),$$

$$F(k) := \mathbb{P}(T_{m+1} - T_m \leq k) = \sum_{t=1}^{k} f(t),$$

and also denote the associated survival/reliability functions respectively by $\overline{F}_{i\bullet}(k)$, $\overline{F}_{\bullet j}(k)$, $\overline{F}(k)$.

2 Estimation of semi-Markov chains

When undertaking estimation of a semi-Markov chain, in practical situations one can have two main types of data: either one sample path or several independent sample paths.

Thus, in the first case, we observe a sample path of a length say n of the semi-Markov chain $(Y_m)_m$, i.e., $y_1, y_2, \ldots, y_n$. Several situations can be encountered here, according to whether the first or the last sojourn times are completely observed or not. To be more specific:

(i) The simplest situation is the one when all the sojourn times are completely observed (non-censored); in other words, in this case, the sample path $y_1, y_2, \ldots, y_n$ can be described as $j_1, k_1, j_2, k_2, \ldots, j_t, k_t, j_{t+1}$, where $j_1, \ldots, j_{t+1}$ are the successive distinct visited states and $k_1, \ldots, k_t$ are the successive sojourn times (observed, non censored);

(ii) A more complicated and realistic framework is when the last sojourn time is not completely observed, thus being right censored. Note that in most practical situations this case occurs;

(iii) An analogous situation is when the first sojourn time is not completely observed, thus being also right censored. In practice, this case occurs when one does not know the beginning of a phenomenon modeled by a semi-Markov chain;

(iv) The most complete framework is when both the first and the last sojourn times are right censored.

In the second case, we observe a certain number, say L, of independent sample paths of the semi-Markov chain $(Y_m)_m$, each of them of length n_l, $l = 1, \ldots, L$, i.e., $y_1^l, y_2^l, \ldots, y_{n_l}^l$. Taking into account the previous discussion, each of these L sample paths can be: a complete sample path (of the type (i) described before), a sample path with the last or the first sojourn time that is right censored (of the types (ii) and (iii) described before), a generally censored

sample path, with the last and the first sojourn times right censored (of the type (iv) described before).

In accordance with what we discussed in the previous section, note also that we can assume that the semi-Markov kernel is of one of the four types we considered: either of the general type $q_{ij}(k) = p_{ij}f_{ij}(k)$, as introduced in (3), or with the holding time distributions $f_{ij}(k)$ depending only on the current state, $q_{ij}(k) := p_{ij}f_{i\bullet}(k)$, as introduced in (5), or with the holding time distributions depending only on the next state to be visited, $q_{ij}(k) := p_{ij}f_{\bullet j}(k)$, as introduced in (6), or with the holding time distributions depending neither on the current nor on the future state, $q_{ij}(k) := p_{ij}f(k)$, as introduced in (7). All these four types of models will yield specific parameter estimators, so, from a practical point of view, is important to study them separately.

Let us consider a parametric estimation framework for the general semi-Markov case, that is for the kernel $q_{ij}(k) = p_{ij}f_{ij}(k)$ (cf. Equation (3)). We are assuming that the holding distributions $f_{ij}(k) = f_{ij}(k;\theta_{ij})$ depend on some unknown parameters $\theta_{ij} \in \mathbb{R}^{m_{ij}}$, where the dimension of the parameter space m_{ij} is known; no assumption is done on $(p_{ij})_{ij}$. Starting from data, we want to estimate p_{ij} and θ_{ij}, $i, j \in E$; thus we have a parametric estimation problem, with the parameter $(\theta_{ij}, p_{ij}, i, j \in E) \in \mathbb{R}^{\sum_{i,j\in E} m_{ij}+s^2}$ (s is the cardinal of E) and constraints $\sum_{j\in E} p_{ij} = 1, i \in E$.

Let us assume that we observe a sample path of a semi-Markov chain of length n censored at the beginning and at the end, i.e.,

$$y_1, y_2, \ldots, y_n,$$

or, equivalently,

$$j_0, k_0, j_1, k_1, j_2, k_2, \ldots, j_t, k_t, j_{t+1}, k_{t+1},$$

with $\sum_{i=0}^{t+1} k_i = n$, where k_0 is the first sojourn time, assumed to be right censored, k_{t+1} is the last sojourn time, assumed to be right censored, while $k_1, \ldots, k_t$ are successive sojourn times, assumed to be complete (observed, non censored).

The *contributions to the likelihood* of the first right censored time k_0 in state j_0 before going to state j_1 and of the last right censored time k_{t+1} in state j_{t+1} have been previously described.

Consequently, the likelihood function is given by

$$
\begin{aligned}
&L(p_{uv}, \theta_{uv}; u, v \in E) \\
&= \mu_{j_0} \prod_{i=0}^{t} p_{j_i j_{i+1}} f_{j_i j_{i+1}}(k_i; \theta_{j_i j_{i+1}}) \overline{F}_{j_0 j_1}(k_0; \theta_{j_0 j_1}) \overline{F}_{j_{t+1}\bullet}(k_{t+1}; \theta_{k_{t+1} v}, v \in E) \\
&= \mu_{j_0} \prod_{u,v \in E} p_{uv}^{N_{uv}(n)} \prod_{u,v \in E} \prod_{k=1}^{n} f_{uv}(k; \theta_{uv})^{N_{uv}(k;n)} \\
&\times \prod_{u \in E} \prod_{k=1}^{n} \overline{F}_{u\bullet}(k; \theta_{uv}, v \in E)^{\mathbb{1}_{\{J_{T_{N(n)}} = u, X_{T_{N(n)+1}} > k\}}} \\
&\times \prod_{u,v \in E} \prod_{k=1}^{n} \overline{F}_{uv}(k; \theta_{uv})^{\mathbb{1}_{\{J_0 = u, J_1 = v, X_1 > k\}}}. \qquad (8)
\end{aligned}
$$

Taking into account the form of the log-likelihood function, we have to consider the cases $j_0 = j_{t+1}$ and $j_0 \neq j_{t+1}$.

Bibliographie

Ayhar, C., Barbu, V. S., Mokhtari, F., & Rahmani, S. (2022). On the Asymptotic Properties of Some Kernel Estimators for Continuous Semi-Markov Processes. *Comm. Statist. Theory Methods*, 34(2), 299–318. https://doi.org/10.1080/10485252.2022.2044033

Barbu, V. S., Lecocq, F., Vergne, N. Parametric estimation of censored semi-Markov chains, In *Stochastic Modeling and Statistical Methods: Advances and Applications*, eds. Alex Karagrigoriou, Sonia Malefaki, Ioannis S. Triantafyllou, Elsevier, 2024

Barbu, V. S., Bérard, C., Cellier, D., Lecocq, F., Lothode, C., Sautreuil, M., & Vergne, N. (2021). R package `smmR` : simulation, estimation and reliability of semi-markov models. https://cran.r-project.org/web/packages/smmR/index.html.

Barbu, V. S., Bérard, C., Cellier, D., Sautreuil, M., & Vergne, N. (2022). SMM: An R Package for Estimation and Simulation of Discrete-Time Semi-Markov Models. *The R Journal*, 10(2), 226–247. https://doi.org/10.32614/RJ-2018-050

Barbu, V. S., Karagrigoriou, A., & Makrides, A. (2016). Semi-Markov Modelling for Multi-State Systems. *Methodology and Computing in Applied Probability*, 18(3), 437–458. https://doi.org/10.1007/s11009-016-9510-y

MAINTENANCES IMPARFAITES AVEC ACTIONS CORRECTIVES DÉCALÉES : MODÉLISATION, INFÉRENCE STATISTIQUE POUR DONNÉES CENSURÉES ET APPLICATION INDUSTRIELLE

Théo Cousino[1] & Florent Brissaud[2] & Leïla Marle[3] & Laurent Doyen[4] & Olivier Gaudoin[5]

[1] *Research & Innovation Center for Energy, GRTgaz, 1 rue du Commandant d'Estienne d'Orves Villeneuve-la-Garenne, 92390, France*
email : theo.cousino@grtgaz.com

[2] *Research & Innovation Center for Energy, GRTgaz, 1 rue du Commandant d'Estienne d'Orves Villeneuve-la-Garenne, 92390, France*
email: florent.brissaud@grtgaz.com

[3] *Research & Innovation Center for Energy, GRTgaz, 1 rue du Commandant d'Estienne d'Orves Villeneuve-la-Garenne, 92390, France*
email: leila.marle@grtgaz.com

[4] *Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 700 AV Centrale Saint-Martin-d'Hères, 38400, France*
email: laurent.doyen@univ-grenoble-alpes.fr

[5] *Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 700 AV Centrale Saint-Martin-d'Hères, 38400, France*
email: olivier.gaudoin@univ-grenoble-alpes.fr

Résumé. On propose un nouveau modèle de maintenance pour un système réparable. Si une défaillance survient entre deux inspections successives, une réparation minimale est effectuée. Au moment de l'inspection, deux types de maintenance préventive peuvent être effectués (MP et MPC). Le choix entre ces deux types de maintenance préventive dépend du fait qu'il y a eu ou non une défaillance depuis l'inspection précédente. Les deux types de maintenance sont supposés imparfaits, du type ARA, avec des paramètres d'efficacité différents.

Nous présentons plusieurs méthodes d'inférence statistique pour estimer les paramètres de ce nouveau modèle de maintenance. Le cas des données complètes est d'abord étudié, puis l'estimation des paramètres se fait pour des données censurées par intervalle, censurées à gauche, pour des données combinant ces deux types de censure, ainsi que pour des données hétérogènes via l'utilisation de covariables. La qualité des méthodes proposées est évaluée sur des données simulées.

Des données réelles, à la fois censurées et hétérogènes, provenant de systèmes passifs, sont étudiées à l'aide des méthodes développées.

Mots-clés. maintenance imparfaite, fiabilité, vieillissement, censure, covariable, estimation statistique

Abstract. A new maintenance model is proposed for a repairable system. If a failure occurs between two successive inspections, a minimum repair is carried out. At the time of the inspection, two types of preventive maintenance can be carried out (PM and PCM). The choice between these two types of

preventive maintenance depends on whether or not there has been a failure since the previous inspection. The two types of maintenance are assumed to be imperfect, of the ARA type, with different efficiency parameters.

We present several statistical inference methods for estimating the parameters of this new maintenance model. The case of complete data is first studied, then the parameters are estimated for interval-censored data, left-censored data, data combining these two types of censoring, and heterogeneous data, through the use of covariates. The quality of the proposed methods is evaluated on simulated data.

Real data, both censored and heterogeneous, from passive system are studied using the methods developed.

Keywords. imperfect maintenance, reliability, aging, censoring, covariate, statistical estimation

1 Maintenances imparfaites avec actions correctives décalées

La fiabilité d'un système dépend des effets conjoints du vieillissement et de l'efficacité des maintenances. Le vieillissement intrinsèque est modélisé à l'aide de lois de probabilité pour la durée de fonctionnement d'un système non maintenu. Parmi ces lois, la plus courante est la loi de Weibull,

La maintenance corrective (MC) survient après une défaillance et a pour but de remettre le matériel en état de fonctionner. La maintenance préventive (MP) est effectuée sur un matériel non défaillant et a pour objectif de ralentir son vieillissement et ainsi de retarder l'occurrence des défaillances. Lorsqu'un système est entretenu, la maintenance peut avoir différents niveaux d'efficacité. Une maintenance parfaite remet le système dans un état neuf, tandis qu'une maintenance minimale ne fait que rétablir son fonctionnement sans ralentir son vieillissement. Entre ces deux extrêmes se trouve la maintenance imparfaite, où le système est rajeuni mais sans retrouver son état d'origine. L'impact de ces maintenances est modélisé à l'aide de modèles d'âge virtuel, en particulier les modèles ARA (Arithmetic Reduction of Age) proposés par Doyen et Gaudoin (2004, 2011), qui ajuste l'âge du système après chaque intervention.

On propose un nouveau modèle de maintenance spécifiquement adapté à notre cadre d'étude. Dans ce cas, chaque défaillance est suivie d'une réparation minimale, et des inspections sont réalisées, au cours desquelles deux types de maintenance préventive peuvent être effectuées (MP et MPC). Si aucune défaillance n'est survenue depuis la dernière inspection, la maintenance préventive, simplement notée MP, correspond à une action de maintenance préventive de base, dont l'effet est planifié par le plan de maintenance initial. Si au moins une défaillance s'est produite depuis la dernière inspection, la maintenance préventive, notée MPC, comprend à la fois l'action préventive prévue par le plan de maintenance initial et une action corrective spécifique qui vise à corriger les défauts survenus depuis la dernière inspection. Les deux types de maintenance sont supposés imparfaits, de type ARA, avec des paramètres d'efficacité différents.

Le nouveau modèle est noté F ABAO - MP M1 - MPC M2 où M1 représente le modèle de maintenance associé aux MPs et M2 le modèle de maintenance associé aux MPCs.

2 Méthodes d'inférence statistique pour données complètes, censurées et hétérogènes

La gestion des actifs industriels nécessite l'application de modèles de maintenance, ainsi que l'estimation statistique de leurs paramètres sur la base de l'observation des données relatives aux défaillances et aux maintenances. Par exemple, Remy et al. (2013) et Guerra de Toledo et al. (2015, 2016) ont réalisé une analyse statistique à partir des modèles ARA sur des données industrielles réelles. Certaines informations sur les données relatives aux défaillances et aux maintenances peuvent être manquantes, c'est ce qu'on appelle la censure. Deux types de censure sont considérés ici :

- censure par intervalle : c'est le cas lorsque les données indiquent qu'un événement s'est produit dans un intervalle de temps spécifique, mais que l'instant exact n'est pas connu.

- censure à gauche : c'est le cas lorsque les données ne sont observées qu'après un certain instant, au-delà de la date de mise en service.

Dans le cas de Processus de Poisson Non Homogène (NHPP), l'estimation des paramètres avec de la censure par intervalle a été étudié dans Peng et al. (2017) et dans da Cunha et al. (2024). De même, la prise en compte de la censure gauche a été abordée dans Dijoux et al. (2016).

On propose ici des méthodes d'inférence statistique pour le nouveau modèle de maintenance. Tout d'abord, le cas des données complètes est examiné. Ensuite, on considère des données censurées. Le cas des données censurées par intervalle est traité dans Cousino et al. (2024). Par la suite, le cas des données censurées à gauche est développé dans Cousino et al. (2025). Différentes méthodes d'estimation sont proposées pour chacun de ces cas. Le cas des données censurées à la fois à gauche et par intervalle est ensuite abordé en tenant compte des différentes méthodes proposées. La qualité des estimations pour chacun des cas est évaluée à partir de données simulées.

Les estimations statistiques tendent souvent à considérer des groupes de systèmes comme homogènes. Cependant, en pratique, les actifs industriels présentent souvent des différences. Il est donc nécessaire de disposer de méthodes capables de prendre en compte cette hétérogénéité. Pour intégrer cette variabilité dans les modèles de maintenance imparfaite, on ajoute des covariables. Les covariables sont fixes et peuvent être quantitatives ou qualitatives.

Les expériences montrent que la qualité des estimations dépend principalement de la taille des données : une bonne qualité d'estimation est obtenue lorsque le nombre de systèmes étudiés et de défaillances est élevé.

3 Application industrielle

On présente une application pratique sur des données réelles provenant de systèmes passifs (S). S est un système de sécurité qui fait partie d'une installation plus large.

Nous considérons dans notre étude 528 S hétérogènes, parmi lesquels 69 ont des données uniquement censurées par intervalles. 459 ont des données censurées à gauche et par intervalle, avec une censure à gauche comprise entre 1 et 5 ans. Les systèmes S présentent des facteurs internes. Des études ont montré que la pression de réglage et l'identifiant du constructeur de S correspondaient aux facteurs internes les plus significatifs. On considère donc à présent une covariable quantitative associée à la pression de réglage de S et une covariable qualitative associée à l'identifiant du constructeur de S.

L'ensemble des paramètres du modèle est noté $\boldsymbol{\theta} = (\eta_1, \eta_2, \eta_3, \beta, \rho_p, \rho_{pc}, \zeta_1)$ où :

- η_1, η_2 et η_3 représentent les différentes modalités de la covariable qualitative associée à η. On fait porter l'effet de cette covariable sur le paramètre d'échelle de la loi de Weibull avec un modèle de type Cox.
- β correspond au paramètre de la loi de Weibull utilisé pour caractériser le vieillissement intrinsèque.
- ρ_p et ρ_{pc} sont les efficacités des MPs et MPCs.
- ζ_1 est le coefficient associé à la covariable quantitative.

Nous considérons le modèle F ABAO- MP ARA$_\infty$ - MPC ARA$_\infty$. On obtient les estimations des paramètres suivants :

F ABAO - MP ARA$_\infty$- MPC ARA$_\infty$						
$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\beta}$	$\hat{\rho}_p$	$\hat{\rho}_{pc}$	$\hat{\zeta}_1$
6588.48	6810.63	5405.89	2.08	0.15	0.89	−0.07

L'estimation de β est proche de 2, ce qui signifie que les S ont un vieillissement significatif. L'estimation de ρ_p égale à 0.15 indique qu'une MP a une efficacité faible. Au contraire, l'estimation de ρ_{pc} égale à 0.89 indique qu'une MPC a une efficacité forte, ce qui est un résultat satisfaisant. L'estimation de ζ_1 proche de 0 indique que la covariable quantitative a très peu d'effet.

4 Conclusion

Un nouveau modèle adapté à notre cadre d'étude a été développé. Celui-ci comprend des réparations minimales et des maintenances préventives imparfaites de type ARA.

Plusieurs méthodes d'inférence statistique pour estimer les paramètres de notre modèle ont été développées. Ces méthodes prennent en compte la censure et l'hétérogénéité dans les données. Ces méthodes ont été appliquées sur des données simulées et réelles. Les estimations sont globalement satisfaisantes lorsque le nombre de données est suffisamment important.

Une perspective de ce travail est de proposer des méthodes bayésiennes pour le cas où très peu de données informatives sont disponibles.

Ces travaux sont les résultats d'une thèse CIFRE de 3 ans entre GRTgaz et le laboratoire Jean Kuntzmann.

Bibliographie

Cousino, T., Brissaud, F., Marle, L., Doyen, L. and Gaudoin, O. (2024). Imperfect maintenance modelling and estimation for interval-censored data. *Reliability Engineering and System Safety*, 252 :110424.

Cousino, T., Brissaud, F., Marle, L., Doyen, L. and Gaudoin, O. (2025). Parametric estimation in an imperfect maintenance model for left censored and interval censored data. En cours de rédaction.

da Cunha et al. (2024). A Bayesian approach for reliability estimation for non-homogeneous and interval-censored failure data. *Process Safety and Environmental Protection*, 182 :775-788.

Dijoux, Y., Fouladirad, M. and Nguyen, D.T. (2016). Statistical inference for imperfect maintenance models with missing data. *Reliability Engineering and System Safety*, 154 : 84-96.

Doyen, L. and Gaudoin, O. (2004). Classes of imperfect repair models based on reduction of failure intensity or virtual age. *Reliability Engineering and System Safety*, 84(1), 45-56.

Doyen, L. and Gaudoin, O. (2011). Modelling and assessment of ageing and efficiency of corrective and planned preventive maintenance. *IEEE Transactions on Reliability*, 60(4), 759-769.

Guerra de Toledo, M., Colosimo, E., Freitas, M. and Gilardoni, G. (2015). ARA and ARI imperfect repair models: Estimation, goodness-of-fit and reliability prediction. *Reliability Engineering and System Safety*, 140, 107-11.

Guerra de Toledo, M. Colosimo, E., Freitas, M., and Gilardoni, G. (2016). Optimal periodic maintenance policy under imperfect repair: A case study on the engines of off-road vehicles. *IIE Transactions*, 48(8), 747-758.

Peng, E., Wang, Y., Zi, Y., Tsui, K. and Zhang, C. (2017). Dynamic reliability assessment and prediction for repairable systems with interval-censored data. *Reliability Engineering & System Safety*, 159 :301-309.

Remy, E., Corset, F., Despréaux, S., Doyen, L. and Gaudoin, O. (2013). An example of integrated approach to technical and economic optimization of maintenance. *Reliability Engineering and System Safety*, 116, 8-19.

Optimisation de la vitesse des trains sous contraintes de ponctualité et de consommation énergétique

Christine Funfschilling [1] & Guillaume Perrin [2]

[1] *SNCF, France, christine.funfschilling@sncf.fr*
[2] *Université Gustave Eiffel, France, guillaume.perrin@univ-eiffel.fr*

Résumé. Ce travail propose une méthode pour optimiser une fonction sous contraintes linéaires et non-linéaires. La fonction recherchée est modélisée par un processus stochastique gaussien. Ce processus est alors conditionné afin que les contraintes linéaires sur le champ et sa dérivée soient respectées. Le problème d'optimisation est alors ramené à la recherche de la meilleure trajectoire de ce processus vis-à-vis des contraintes non-linéaires et d'un critère de performance intégré. Les développements sont appliqués à l'optimisation des trajectoires de vitesse de trains vis-à-vis de critères énergétiques, sous des contraintes de sécurité et de confort. Ce document présente d'abord la problématique, la méthode de conditionnement par les contraintes linéaires puis d'adaptation des trajectoires pour vérifier les contraintes non-linéaires et décrit pour finir l'application ferroviaire considérée.

Mots-clés. Conditionnement gaussien adaptatif, contraintes linéaires et non-linéaires.

Abstract. This work proposes a method for optimising a function under linear and non-linear constraints. The desired function is modelled by a Gaussian stochastic process. This process is then conditioned so that the linear constraints on the field and its derivative are respected. The optimisation problem is then reduced to finding the best trajectory for this process with respect to the non-linear constraints and an integrated performance criterion. The developments are applied to the optimisation of train speed trajectories with respect to energy criteria, under safety and comfort constraints.

Keywords. Adaptative Gaussian process regression, linear and non-linear contraints.

Les trajectoires de vitesse des trains sont contraintes pour des raisons de confort et de sécurité (limites sur la vitesse, sur les accélérations latérales et longitudinales), pour des raisons opérationnelles (arrêts en gare, correspondances, croisement des trains) mais aussi pour des raisons physiques puisque les efforts de traction et de freinage sont limités par l'adhérence au contact roue /rail par la chaîne de traction et les systèmes de freinage. On observe cependant des variations significatives en consommation énergétique sur le même trajet. Celles-ci peuvent-être attribuées aux modifications environnementales, aux caractéristiques du train mais aussi à la commande du conducteur. L'objectif de ce travail est de proposer une méthode de construction de trajectoires de vitesse qui respectent ces contraintes en limitant la consommation énergétique.

1 Définition du problème

Dans ce travail on cherche à optimiser le comportement dynamique non-linéaire d'un système décrit par l'équation différentielle paramétrique suivante :

$$\ddot{x}(t) = f(x(t), \dot{x}(t); \boldsymbol{z_f}), \quad t \in \mathbb{R}^+, \tag{1}$$

avec $\boldsymbol{z_f}$ les paramètres caractérisant le train. Notons alors $t \mapsto x(t; \boldsymbol{z}_f)$ la fonction solution de ce système associée au vecteur $\boldsymbol{z_f}$. Cette fonction solution doit par ailleurs vérifier les contraintes suivantes :

$$\begin{cases} x_i \leq x(t; \boldsymbol{z}_f) \leq x_{i+1}, \forall t \in [t_i^\ell, t_i^u], i \in [\![1, N]\!], \\ 0 \leq \dot{x}(\cdot; \boldsymbol{z}_f) \leq u_1, \\ \ddot{x}(\cdot; \boldsymbol{z}_f) \leq u_2(x(\cdot; \boldsymbol{z}_f), \dot{x}(\cdot; \boldsymbol{z}_f)). \end{cases} \tag{2}$$

Les fonctions u_1 et u_2 sont supposées connues, et nous appellerons $\mathcal{E}(\boldsymbol{z}_f)$ l'ensemble des solutions de (1) qui vérifient ces contraintes. En particulier, on se concentrera sur des trajectoires devant se trouver en x_f à l'instant T_f (contrainte de position finale et de temps final), ce qui se traduit par $t_N^u = t_N^\ell = T_f$ et $x_N = x_{N+1} = x_f$.

La performance de ce système est décrite par une autre fonction h (également supposée connue), paramétrée par un autre vecteur de paramètres $\boldsymbol{z}_h$, qui caractérise la consommation énergétique. Le problème de minimisation s'écrit alors :

$$\min_{x \in \mathcal{E}(\boldsymbol{z}_f)} \int_0^{T_f} h(x(t; \boldsymbol{z}_f); \boldsymbol{z}_h)\ dt. \tag{3}$$

2 Conditionnement gaussien adaptatif

Pour résoudre le problème d'optimisation défini par l'Eq. (3), on commence par modéliser la solution du problème comme une réalisation particulière d'un processus Gaussien: $X \sim \mathrm{GP}(\mu, C)$, avec μ et C les fonctions moyenne et covariance de X. Les paramètres caractérisant ces fonctions μ et C sont choisis de manière à respecter l'ensemble des contraintes linéaires, voir Perrin et al. (2021). La trajectoire optimale est alors recherchée de manière séquentielle, en découpant $[0, T_f]$ en M intervalles successifs $[t_i, t_{i+1}]$, avec $t_0 = 0$ et $t_M = T_f$. Sur chaque intervalle $[t_i, t_{i+1}]$, en partant de $i = 0$, des trajectoires possibles vis-à-vis des contraintes linéaires (et potentiellement vis-à-vis de nouvelles contraintes de continuité) sont générées, et celle qui minimise l'intégrale de h entre 0 et t_{i+1} parmi ces dernières est sélectionnée. Pour finir, la trajectoire ainsi reconstruite sur tout $[0, T_f]$ est localement adaptée pour respecter les contraintes non-linéaires.

3 Application ferroviaire

Une application à la construction d'une trajectoire de vitesse optimisée pour une ligne à grande vitesse est finalement présentée. La dynamique du train est modélisée en considérant

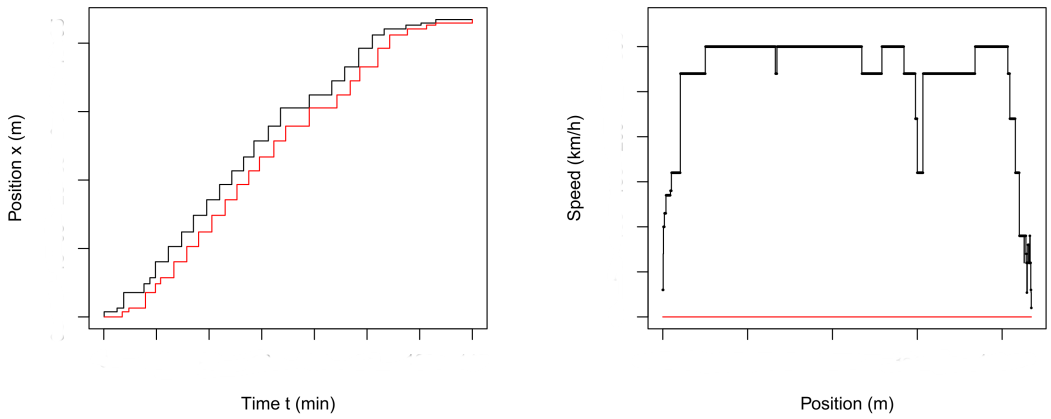


Figure 1: Contraintes sur la position (gauche) et sur la vitesse (droite).

les corps rigides et en intégrant les termes de frottement, de traction, de freinage et de poids. La puissance consommée est alors estimée à partir des efforts de traction et de freinage. Les contraintes linéaires sont données à la figure 3.

Un exemple de trajectoire de vitesse optimisée est donné à la figure 2.

Conclusions

Une procédure permettant de construire des trajectoires optimales vis-à-vis d'un critère intégré et sous contraintes linéaires et non-linéaires a été mise en place. La trajectoire présentant la meilleure performance est finalement choisie dans un ensemble de trajectoires générées et est localement adaptée pour vérifier les contraintes non-linéaires.

Bibliographie

Perrin G. and Da Veiga S. (2021), Constrained Gaussian process regression: an adaptive approach for the estimation of hyperparameters and the verification of constraints with high probability, *Journal of Machine Learning for Modeling and Computing*.

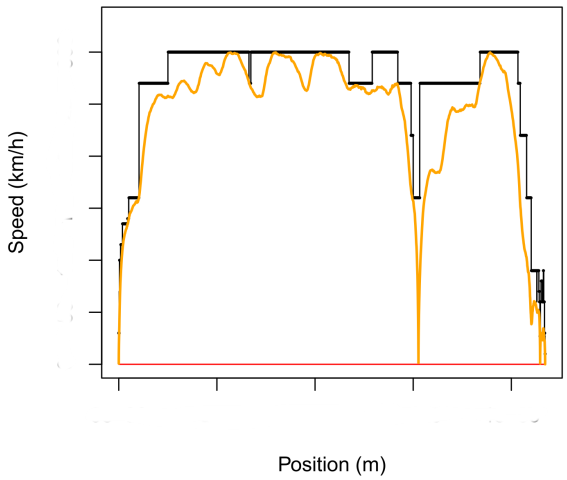


Figure 2: Exemple de vitesse optimisée (en orange) respectant les contraintes linéaires et non-linéaires.

Étude théorique et empirique de quelques mesures de risque pour les systèmes industriels critiques

Marie Temple-Boyer[1,2] & Guillaume Perrin [2] & Vincent Chabridon [1] & Julien Pelamatti [1] & Emmanuel Remy [1] & Emma Cussenot [3]

[1] *EDF R&D, 6 quai Watier, 78401 Chatou, France, (first.last@edf.fr)*
[2] *Université Gustave Eiffel, COSYS, 14-20 Boulevard Newton, 77447 Marne-la-Vallée, France, (first.last@univ-eiffel.fr)*
[3] *CentraleSupélec, 3 rue Joliot Curie, 91190 Gif-sur-Yvette, France, (first.last@student-cs.fr)*

Résumé. L'évaluation de la fiabilité de systèmes industriels critiques (tels que les systèmes de production d'électricité) s'appuie sur une analyse d'incertitudes et, in fine, sur l'estimation d'une « mesure de risque » qui caractérise le risque de défaillance du système. Traditionnellement, les mesures de risque rencontrées en pratique correspondent généralement à des quantiles d'ordre élevé ou à des probabilités de dépassement de seuil d'une variable d'intérêt caractérisant le comportement du système. Étant donné le caractère généralement « fiable » de tels systèmes critiques, le nombre de défaillances observées est très faible, induisant ce que l'on appelle communément un problème d'estimation d'événements rares. Toutefois, un plus large panel de mesures de risque peut être considéré en pratique. En effet, les quantités telles que la marge, le superquantile, ou bien la notion de probabilité de défaillance « amortie » sont aussi des mesures de risque qui peuvent s'avérer pertinentes pour les analyses de risques en ingénierie. Ainsi, le but de cette contribution est double : (i) d'une part, elle s'attelle à présenter les différentes mesures de risque et leurs propriétés théoriques afin de juger de leur pertinence au regard d'une prise de décision dans un contexte industriel ; (ii) d'autre part, un focus est porté sur le comportement numérique d'estimateurs du superquantile et de la probabilité de défaillance amortie, ces quantités étant moins utilisées en pratique, tout en offrant des propriétés pertinentes vis-à-vis d'autres mesures de risque. Ces travaux préliminaires visent à établir une démarche méthodologique cohérente dans l'estimation de mesures de risque et de formalisation future de critères de décision avec garanties.

Mots-clés. Quantification d'incertitudes, mesures de risque, quantile, superquantile, probabilité de défaillance, probabilité de défaillance amortie

Abstract. Reliability assessment of critical industrial systems (such as power generation systems) is based on an analysis of uncertainties and, ultimately, on the estimation of a "risk measure" characterizing the risk of a system failure. Traditionally, the risk measures encountered generally correspond to high-order quantiles or probabilities of exceeding a threshold of a variable of interest characterizing the system's behavior. Given the generally "reliable" nature of such critical systems, the number of observed failures is very low, leading to what is commonly referred to as a rare-event estimation problem. However, a wider range of risk measures can be considered in practice. Indeed, quantities such as the margin,

the superquantile, or the notion of "buffered failure probability" are also risk measures that may be relevant to engineering risk analysis. Thus, the aim of this contribution is twofold: (i) on the one hand, it recalls the definition of various risk measures and their theoretical properties while highlighting their relevance to decision-making in an industrial risk analysis context; (ii) on the other hand, it focuses on the numerical behavior of some estimators of the superquantile and the buffered failure probability, these quantities being less used in practice but presenting relevant properties. The aim of this preliminary work is to establish a consistent methodological approach to estimating risk measures and formalizing decision criteria with guarantees.

Keywords. Uncertainty quantification, risk measures, quantile, superquantile, failure probability, buffered failure probability

1 Introduction

Pour assurer la fiabilité des systèmes critiques de production électrique, EDF utilise des outils de simulation numérique couplés à des analyses d'incertitudes dans le but de maîtriser les marges de sûreté exigées par les autorités de réglementation. Ces systèmes critiques étant soumis à de nombreuses sources d'incertitudes (p.ex., variabilité intrinsèque, incertitudes statistiques, méconnaissances, erreurs de modèle), une démarche méthodologique générique, commune à de nombreux acteurs industriels et académiques, est généralement mise en place [Blanchard et al., 2023]. Au sein de cette démarche générique, l'analyse de fiabilité consiste généralement à caractériser le comportement des systèmes dans des régimes rares de défaillance entraînant des conséquences indésirables, tant sur les volets humains, environnementaux, sociétaux ou financiers. Selon les domaines d'application, les pratiques et critères utilisés dans l'analyse de fiabilité varient. À titre d'exemple, on peut mentionner, d'une part, les analyses menées en sûreté nucléaire pour lesquelles on estime un quantile d'une variable aléatoire d'intérêt (par exemple une température). D'autre part, dans le domaine hydroélectrique, la fiabilité des conduites forcées est généralement étudiée à travers le calcul d'une probabilité de défaillance. Enfin, en analyse des risques financiers, ce sont principalement des quantiles et superquantiles qui sont classiquement considérés. Toutes les grandeurs précédemment mentionnées sont appelées des *mesures de risque*. Dans cette communication, une étude théorique et empirique de quelques unes de ces mesures de risque est proposée, avec un attention particulière pour le *superquantile* et la *probabilité de défaillance* dite *amortie* [Rockafellar and Royset, 2015].

Dans ce qui suit, on se donne une variable d'intérêt scalaire Y, que l'on considère aléatoire à densité continue. On suppose aussi que cette variable Y correspond à un marqueur du niveau de fiabilité du système dans le sens où, plus Y est élevée, moins le système est fiable.

Dans une première partie, la notion de mesure de risque est définie et on rappelle une liste de potentielles propriétés « souhaitables » pour ces mesures de risque. À l'aune de ces propriétés, deux catégories de mesures de risque peuvent être distinguées : celles qui sont homogènes à la quantité d'intérêt Y (marge, quantile, superquantile) et celles qui sont homogènes à une probabilité (probabilité de défaillance, probabilité de défaillance amortie). Une

telle distinction joue un rôle, notamment, dans l'évaluation de la pertinence des propriétés listées précédemment. Dans une deuxième partie, nous nous focalisons sur le superquantile et la probabilité de défaillance amortie et proposons d'étudier, sur une fonction analytique, le comportement numérique de quelques estimateurs.

2 Mesures de risque : définition et propriétés

Une *mesure de risque* est une application qui, à une variable Y, associe un scalaire $\mathcal{R}_Y$. Un tel objet a pour but de donner une quantité résumée du risque associé à la variable aléatoire Y. Si on appelle E un sous-espace vectoriel des variables aléatoires réelles à densité, alors, une mesure de risque sur E est une application $\mathcal{R} : E \to \mathbb{R}, Y \mapsto \mathcal{R}_Y$.

Différentes mesures de risque sont proposées dans la littérature. Certaines sont homogènes en la variable d'intérêt Y, comme, par exemple : (i) la marge $\mathbb{E}(Y) + \lambda\sigma(Y)$ pour $\lambda > 0$ où $\mathbb{E}(Y)$ et $\sigma(Y)$ sont respectivement l'espérance et l'écart-type de la variable aléatoire Y ; (ii) q_α le quantile d'ordre α (pour $0 < \alpha < 1$) défini par $q_\alpha := F_Y^{-1}(\alpha)$ si la fonction de répartition F_Y est inversible ou, plus généralement, par $q_\alpha := \inf\{y \in \mathbb{R} \mid F_Y(y) \geq \alpha\}$; (iii) le superquantile $\overline{q}_\alpha := \mathbb{E}(Y \,|\, Y \geq q_\alpha)$. D'autres sont homogènes à une probabilité, i.e., à valeur dans $[0, 1]$, comme (iv) la probabilité de défaillance $p(s) := \mathbb{P}(Y > s)$ avec $s \in \mathbb{R}$ un seuil ; (v) la probabilité de défaillance amortie $\overline{p}(s)$, définie par $\overline{p}(s) := 1 - \overline{q}_s^{-1}$ où $s \mapsto \overline{q}_s^{-1}$ est la bijection réciproque du superquantile $\alpha \mapsto \overline{q}_\alpha$. Autrement dit, pour $s \in \mathbb{R}$, $\overline{p}(s) = 1 - \alpha$ où α est tel que $\overline{q}_\alpha = s$. Outre ce panel, la littérature s'intéresse à la caractérisation de propriétés « souhaitables » pour une mesure de risque [Artzner et al., 1999, Rockafellar and Royset, 2015].

Définition 1 (Propriétés d'une mesure de risque)**.** *Soit $\mathcal{R}$ une mesure de risque sur E, sous-espace vectoriel des variables aléatoires réelles. On définit les propriétés suivantes :*

1. *(Monotonie). Si $(Y_1, Y_2) \in E^2$ sont tels que $Y_1 \leq Y_2$ presque sûrement (p.s.), alors $\mathcal{R}_{Y_1} \leq \mathcal{R}_{Y_2}$.*
2. *(Continuité). L'application $\mathcal{R} : E \to \mathbb{R}$ est continue pour un certain type de convergence sur E.*
3. *(Homogénéité positive). Pour tout $Y \in E$ et tout $\lambda > 0$, on a $\mathcal{R}_{\lambda Y} = \lambda \mathcal{R}_Y$.*
4. *(Invariance par translation). Pour tout $Y \in E$ et tout $a \in \mathbb{R}$, on a $\mathcal{R}_{Y+a} = \mathcal{R}_Y + a$.*
5. *(Équivalence des constantes). Pour tout $Y \in E$ et $c \in \mathbb{R}$ telles que $Y = c$ p.s., alors $\mathcal{R}_Y = c$.*
6. *(Aversion au risque). Pour tout $Y \in E$ non constante p.s., on a $\mathcal{R}_Y > \mathbb{E}(Y)$.*
7. *(Sous-additivité). Pour tout $(Y_1, Y_2) \in E^2$, on a $\mathcal{R}_{Y_1+Y_2} \leq \mathcal{R}_{Y_1} + \mathcal{R}_{Y_2}$.*
8. *(Convexité). L'application $\mathcal{R} : E \to \mathbb{R}$ est convexe.*
9. *(Fermeture). Pour tout $c \in \mathbb{R}$, l'ensemble $\{Y \in E \mid \mathcal{R}_Y \leq c\}$ est fermé dans E pour un certain type de convergence sur E.*

La formulation mathématique de ces propriétés appelle quelques remarques. En effet, la *monotonie* est une propriété conforme à l'intuition d'une mesure de risque : si une variable

aléatoire Y_2 est associée à une situation moins fiable que Y_1, alors la mesure de risque doit refléter cette relation. Quant à la *continuité*, elle traduit une idée de régularité : deux mesures de risque sont proches si les deux variables aléatoires sous-jacentes le sont suffisamment. L'*homogénéité positive* témoigne du caractère homogène de la mesure de risque en la variable aléatoire associée, et garantit, entre autres, qu'un changement d'unité ne changerait pas les conclusions auxquelles on pourrait aboutir. L'*invariance par translation* témoigne aussi d'une homogénéité de la mesure de risque en la variable aléatoire associée, en garantissant le fait qu'ajouter une constante ne fait qu'ajouter cette même constante à la mesure de risque. L'*équivalence des constantes* rend compte du fait qu'une mesure de risque n'affecte pas la nature d'une variable aléatoire presque sûrement constante. L'*aversion au risque* permet de s'assurer que la mesure de risque considérée est toujours plus pénalisante que l'espérance. La *sous-additivité* et la *convexité* sont équivalentes si la mesure de risque est positivement homogène. La convexité est une propriété pratique si l'on s'intéresse à l'optimisation d'une mesure de risque. Enfin, la *fermeture* joue le rôle d'une hypothèse technique utilisée en optimisation (voir [Rockafellar and Royset, 2015]).

La monotonie est une propriété que l'on considérera prépondérante. En effet, l'utilisation d'une mesure de risque non monotone pourrait aboutir à privilégier, parmi deux situations, une qui est moins fiable. Le quantile, le superquantile et la probabilité de défaillance (amortie ou non) vérifient cette propriété. Dans une moindre mesure, la continuité est aussi une propriété importante : elle permet une régularité qui garantit que la mesure de risque ne va pas varier abruptement si l'on modifie peu la variable aléatoire correspondante.

Les mesures de risque citées précédemment peuvent être regroupées en deux catégories : (i) celles qui sont homogènes en la variable d'intérêt Y (marge, quantile, superquantile) ; (ii) celles qui sont homogènes à une probabilité (probabilité de défaillance, probabilité de défaillance amortie). Pour ces deux dernières, les propriétés 3 à 6 ne sont pas vérifiées. Plus généralement, ces propriétés ne peuvent pas être vérifiées par des mesures de risque homogènes à une probabilité, non pas à cause d'une lacune de ces mesures, mais simplement car elles sont à valeurs dans $[0, 1]$. Aussi, l'homogénéité positive, l'équivalence des constante, l'invariance par translation et l'aversion au risque ne sont pas de propriétés pertinentes à considérer pour des mesures de risque homogènes à une probabilité. Il est alors pertinent de faire une distinction entre les deux types de mesures de risque (i) et (ii) énoncées précédemment et ce notamment en vue de leur utilisation dans des analyses de fiabilité industrielle.

Si le quantile, la probabilité de défaillance et la marge sont très largement rencontrées en pratique dans les analyses de de risque et de fiabilité industrielles, le superquantile et la probabilité de défaillance amortie sont, quant à elles, des mesures beaucoup moins usitées, et sont toujours sujettes à des travaux de recherche actifs.

3 Estimation du superquantile et de la probabilité de défaillance amortie

En pratique, on ne connaît généralement pas la loi de la variable d'intérêt Y, mais on a seulement accès à des réalisations de cette variable aléatoire. Dans cette communication,

nous proposons donc l'étude de deux estimateurs du superquantile et d'un estimateur de la probabilité de défaillance amortie. Plus précisément, on considère un échantillon $(Y_1, \ldots, Y_n)$ indépendant et identiquement distribué selon la loi de Y. On note la statistique d'ordre $(Y_1^* \leq Y_2^* \leq \cdots \leq Y_n^*)$ associée à l'échantillon $(Y_1, \ldots, Y_n)$. Pour $x \in \mathbb{R}$, on note $\lfloor x \rfloor$ la partie entière inférieure de x et $[x]_+ := \max(x, 0)$ la partie positive de x.

Soit $0 < \alpha < 1$ un niveau de confiance, on considère les estimateurs du superquantile $\widehat{\overline{q}}_{\alpha,n}^{\text{Lab}}$ [Labopin-Richard, 2016] et $\widehat{\overline{q}}_{\alpha,n}^{\min}$ [Mafusalov et al., 2018] définis comme suit :

$$\widehat{\overline{q}}_{\alpha,n}^{\text{Lab}} := \frac{1}{(1-\alpha)n} \sum_{i=\lfloor n\alpha \rfloor + 1}^{n} Y_i^* \quad \text{et} \quad \widehat{\overline{q}}_{\alpha,n}^{\min} := \min_{c \in \mathbb{R}} \left(c + \frac{1}{1-\alpha} \frac{1}{n} \sum_{i=1}^{n} [Y_i - c]_+ \right).$$

Soit $s \in \mathbb{R}$ un seuil, l'estimateur de la probabilité de défaillance amortie $\widehat{\overline{p}}_n(s)$ est alors donné par [Mafusalov et al., 2018] :

$$\widehat{\overline{p}}_n(s) := \min_{a \geq 0} \left(\frac{1}{n} \sum_{i=1}^{n} [a(Y_j - s) + 1]_+ \right) \text{ si } s < \max(Y_1, ..., Y_n) \text{ et } 0 \text{ sinon.}$$

Dans ce travail, on se propose d'étudier le comportement de ces estimateurs en faisant varier la taille n de l'échantillon et la loi de Y, notamment en considérant des lois à queue lourde. Pour illustrer ces analyses, la Figure 1 présente une étude des deux estimateurs du superquantile sur le cas où Y suit une loi normale centrée réduite. On constate que, sauf pour une petite taille d'échantillon ($n = 50$), les deux estimateurs se comportent dans ce cas de la même manière.

4 Conclusions et perspectives

Pour conclure, on peut rappeler que la notion de mesure de risque permet d'obtenir une valeur scalaire associée à une variable aléatoire témoignant d'un niveau de fiabilité du système considéré. De nombreuses mesures de risque existent dans la littérature et remplissent certaines propriétés qu'il convient de discuter, notamment à l'aune de leur nature : (i) celles qui sont homogènes en Y (quantile, superquantile, marge), pour lesquelles toutes les propriétés de la Définition 1 sont pertinentes à étudier ; (ii) et celles qui sont homogènes à une probabilité (probabilité de défaillance, probabilité de défaillance amortie), pour lesquelles les propriétés 3 à 6 ne sont pas pertinentes. En perspective de ces travaux, plusieurs pistes sont envisagées : d'une part, approfondir l'analyse théorique des liens qui existent entre les mesures de risque homogènes à Y et celles qui sont homogènes à une probabilité ; d'autre part, étoffer l'étude théorique (dans le cas asymptotique et en n fini) et numérique d'estimateurs associés à toutes les mesures de risque mentionnées ; enfin, étudier, dans l'esprit de [Perrin et al., 2024], le lien avec un cadre décisionnel permettant de constuire un critère de majoration de l'estimateur qui garantisse, avec une probabilité élevée, que l'on ait bien $\mathcal{R}_Y \leq s$.

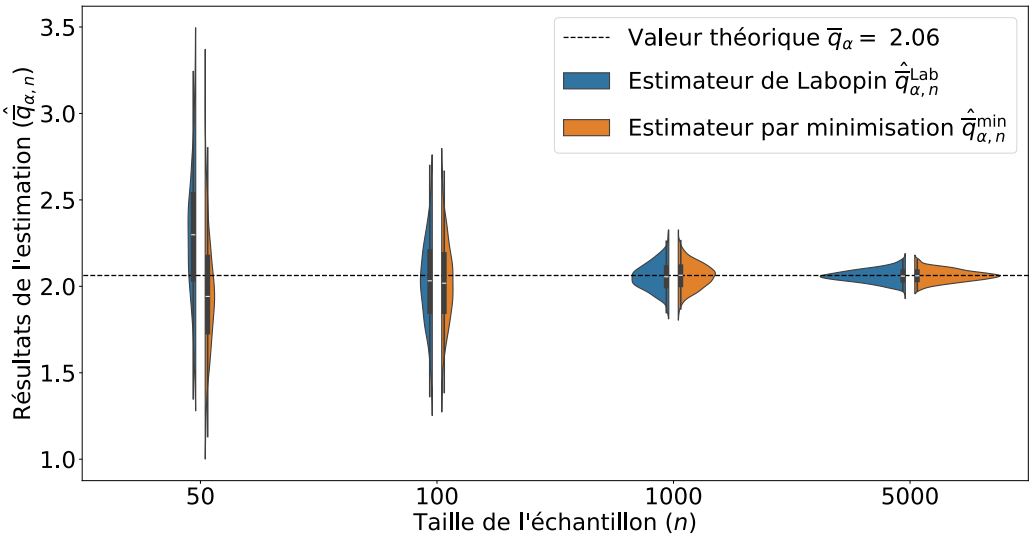


FIGURE 1 – Graphes en violon associés aux deux estimateurs du superquantile $\overline{q}_\alpha$ pour une loi normale centrée réduite avec choix de $\alpha = 0.95$ en fonction de $n \in \{50, 100, 1000, 5000\}$ et considérant 10^3 répétitions.

Références

[Artzner et al., 1999] Artzner, P., Delbaen, F., Eber, J., and Heath, D. (1999). Coherent Measures of Risk. *Mathematical Finance*, 9(3) :203–228.

[Blanchard et al., 2023] Blanchard, J.-B., Chocat, R., Damblin, G., Baudin, M., Bousquet, N., Chabridon, V., Iooss, B., Keller, M., Pelamatti, J., and Sueur, R. (2023). Fiches pédagogiques sur le traitement des incertitudes dans les codes de calcul (I3P : EDF - CEA - Framatome). *hal-04205632*, pages 1–36.

[Labopin-Richard, 2016] Labopin-Richard, T. (2016). *Méthodes statistiques et d'apprentissage pour l'estimation de quantiles et de superquantiles dans des modèles de codes numériques ou stochastiques*. PhD thesis, Université Toulouse 3 Paul Sabatier.

[Mafusalov et al., 2018] Mafusalov, A., Shapiro, A., and Uryasev, S. (2018). Estimation and asymptotics for buffered probability of exceedance. *European Journal of Operational Research*, 270(3) :826–836.

[Perrin et al., 2024] Perrin, G., Reygner, J., and Chabridon, V. (2024). Enhancing Reliability Analysis with Limited Observations : A Statistical Framework for System Safety Margins.

[Rockafellar and Royset, 2015] Rockafellar, R. and Royset, J. (2015). Engineering Decisions under Risk Averseness. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A : Civil Engineering*, 1(2).

Environnement et statistique

Un modèle d'EDS contraint pour l'analyse de trajectoires de narvals

Alexandre Delporte [1] & Adeline Samson [2]

[1] *Université Grenoble-Alpes, France, alexandre.delporte@univ-grenoble-alpes.fr*
[2] *Université Grenoble-Alpes, France, adeline.leclercq-samson@univ-grenoble-alpes.fr*

Résumé. Les narvals en Arctique sont de plus en plus exposés aux activités humaines qui peuvent menacer leur survie temporairement ou de manière permanente en modifiant leur comportement. Nous examinons les données GPS d'une population de narvals exposés au bruit de navires et de canons à air sismiques lors d'une expérience d'exposition contrôlée en 2018 dans le système de fjords de Scoresby Sound au sud-est du Groenland. Le système de fjords a un littoral complexe, limitant les options de réponse comportementale permettant aux narvals d'échapper aux menaces. Nous proposons un nouveau modèle de vitesse corrélée en temps continu avec des coefficients variables qui inclut des contraintes spatiales sur le mouvement. Pour évaluer l'effet de l'exposition au son, nous comparons un modèle de base pour le mouvement avant l'exposition à un modèle de réponse pour le mouvement pendant l'exposition. Notre modèle, appliqué aux données réelles, indique que l'exposition au son peut perturber le mouvement du narval jusqu'à quelques dizaines de kilomètres. Plus précisément, nous avons constaté une augmentation de la vitesse et une diminution de la persistance du mouvement.

Keywords. Ecologie du mouvement, equations différentielles stochastiques, contrainte spatiale, réponse comportementale, modèle à espace d'états

Abstract.

Narwhals in the Arctic are increasingly exposed to human activities that can temporarily or permanently threaten their survival by modifying their behavior. We examine GPS data from a population of narwhals exposed to ship and seismic airgun noise during a controlled experiment in 2018 in the Scoresby Sound fjord system in Southeast Greenland. The fjord system has a complex shore line, restricting the behavioral response options for the narwhals to escape the threats. We propose a new continuous-time correlated velocity model with varying coefficients that includes spatial constraints on movement. To assess the sound exposure effect we compare a baseline model for the movement before exposure to a response model for the movement during exposure. Our model, applied to the narwhal data, indicates that sound exposure can disturb narwhal motion up to a couple of tens of kilometers. Specifically, we found an increase in velocity and a decrease in the movement persistence.

Keywords. Movement ecology, stochastic differential equations, spatial constraint, behavioral response, state space model

1 Données trajectorielles

Six narvals ont été équipés de récepteurs GPS FastLoc en août 2018 dans le Scoresby Sound, au sud-est du Groenland. Un navire militaire de patrouille en mer a été utilisé pour tirer des canons à air sous l'eau entre le 25 août et le 1er septembre. Les données incluent la latitude et la longitude des narvals, la distance relative au navire en mètres, le temps GPS, et la distance à la côte en mètres. Les positions GPS sont connues uniquement lorsque les narvals sont en surface. L'intervalle de temps médian entre deux mesures GPS dans les données est d'environ 5 minutes. Alors que l'analyse statistique dans l'article de Heide-Jorgensen et al. (2021) s'est appuyée sur des données augmentées avec des positions interpolées linéairement chaque seconde, nous considérons ici uniquement les mesures GPS réelles avec des intervalles de temps irréguliers. Les trajectoires GPS sont affichées en Figure 1. Nous définissons la covariable d'exposition E^{ship} comme l'inverse de la distance au navire lorsque le narval est en ligne de vue avec celui-ci.

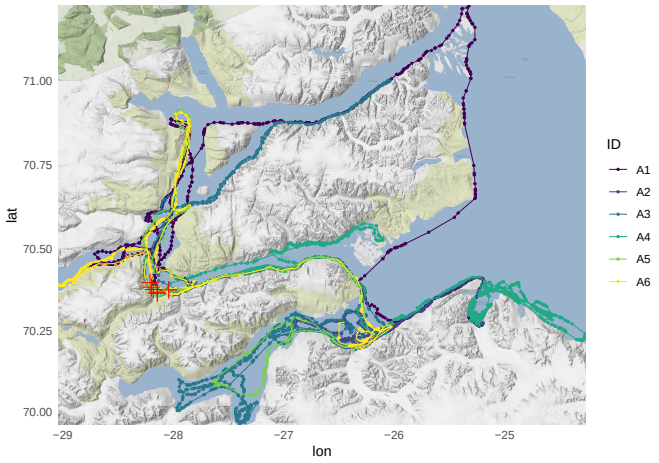


Figure 1: Trajectoires GPS observées. Les croix rouges indiquent les points de départ des trajectoires.

2 Modèle de diffusion avec contrainte spatiale

Nous utilisons un processus de type CVM (Correlated Velocity Model) défini dans l'article de Gurarie (2017). Le nouveau modèle repose sur un paramètre de tortuosité, ou vitesse angulaire ω, pour décrire comment les animaux tournent en réaction au bord. Plus précisément, le processus position X est solution de l'EDS suivante :

$$\begin{cases} dX(t) = V(t)dt \\ dV(t) = -A(X(t), V(t))V(t)dt + \frac{2\nu}{\sqrt{\pi\tau}}dW(t) \end{cases} \tag{1}$$

avec

$$A(X(t), V(t)) = \begin{pmatrix} \frac{1}{\tau} & -\omega(X(t), V(t)) \\ \omega(X(t), V(t)) & \frac{1}{\tau} \end{pmatrix}$$

$V(t)$ est la vitesse horizontale au temps t, $X(t)$ est la position, typiquement en longitude et latitude ou en coordonnées UTM, $W(t)$ est un mouvement brownien bidimensionnel. Le paramètre τ est un temps d'autocorrélation, ν contrôle la norme de la vitesse et la variabilité aléatoire, et ω est une vitesse angulaire qui contrôle la rotation du vecteur vitesse. $\omega(X(t), V(t)) = f_\omega(D^{shore}(t), \Theta(t))$ avec $D^{shore}(t)$ la distance au bord, Θ l'angle entre la vitesse et le vecteur normal au bord et f_ω une fonction lisse de $\mathbb{R}^+ \times]-\pi, \pi]$ vers $\mathbb{R}$.

Nous proposons un modèle paramétrique pour la fonction f_ω. Pour $\Theta \in]-\pi, \pi]$ et $D^{shore} \geq 0$, nous définissons f_ω comme la somme d'un terme répulsif et d'un terme attractif.

$$f_\omega(\Theta, D^{shore}) = g_\omega(\Theta, D^{shore}) + h_\omega(\Theta, D^{shore})) \tag{2}$$

où

$$g_\omega(\Theta, D^{shore}) = a\Theta\left(\Theta - \frac{\pi}{2}\right)\left(\Theta + \frac{\pi}{2}\right) \times \frac{1}{D_{shore}} \exp\left(-\frac{D_{shore}}{D_0}\right)$$

est le terme répulsif et

$$h_\omega(\Theta, D^{shore}) = b \exp\left(-\left(\begin{pmatrix}\Theta & D^{shore}\end{pmatrix} - m\right)\Sigma^{-1}\left(\begin{pmatrix}\Theta & D^{shore}\end{pmatrix} - m\right)^\top\right)$$

est le terme attractif. Ici, $m = (\frac{\pi}{2\sqrt{3}}, D_1)$, $\Sigma = \begin{pmatrix}\sigma_D^2 & 0 \\ 0 & \sigma_\Theta^2\end{pmatrix}$ et $D_0, D_1 > 0$, $a, b > 0$, $\sigma_D, \sigma_\Theta > 0$ sont des paramètres. Les paramètres a et b contrôlent l'amplitude de la vitesse angulaire, D_0 contrôle la distance à laquelle la vitesse commence à tourner pour modéliser la répulsion au bord du domaine, et D_1 contrôle la distance à laquelle la vitesse commence à tourner pour modéliser l'attraction vers le bord. Ensuite, σ_D et σ_θ sont les écarts types dans les noyaux gaussiens.

On obtient une solution approximée en choisissant un petit pas de temps Δ et en approchant D^{shore} et Θ par des fonctions constantes par morceaux à chaque pas de temps. Nous utilisons la densité de transition explicite obtenue en supposant ω constant sur chaque pas de temps. Les trajectoires simulées dans un système de fjords au sud-est du Groenland sont illustrées dans la Figure 2 et comparées à des trajectoires simulées à partir d'un modèle d'EDS réfléchie parfois utilisé en écologie du mouvement, notamment dans le papier de Hanks et al. (2017).

3 Modèle à effets mixtes

Tout d'abord, un modèle de référence est ajusté uniquement sur les données collectées avant l'exposition, afin d'obtenir une estimation du comportement naturel des narvals, lorsqu'ils ne sont pas exposés aux sons du navire. Ensuite, un modèle de réponse est ajusté sur les données collectées pendant l'exposition pour tester si les paramètres estimés s'écartent significativement des valeurs de référence. Pour le modèle de référence, seul l'effet de la côte est inclus via les covariables D^{shore} et Θ. Pour chaque narval $i \in 1, \cdots, N$ et chaque instant $j \in 1, \cdots, n_{i,pre}$, y_{ij} est une observation avec erreur de mesure de la position X_{ij} :

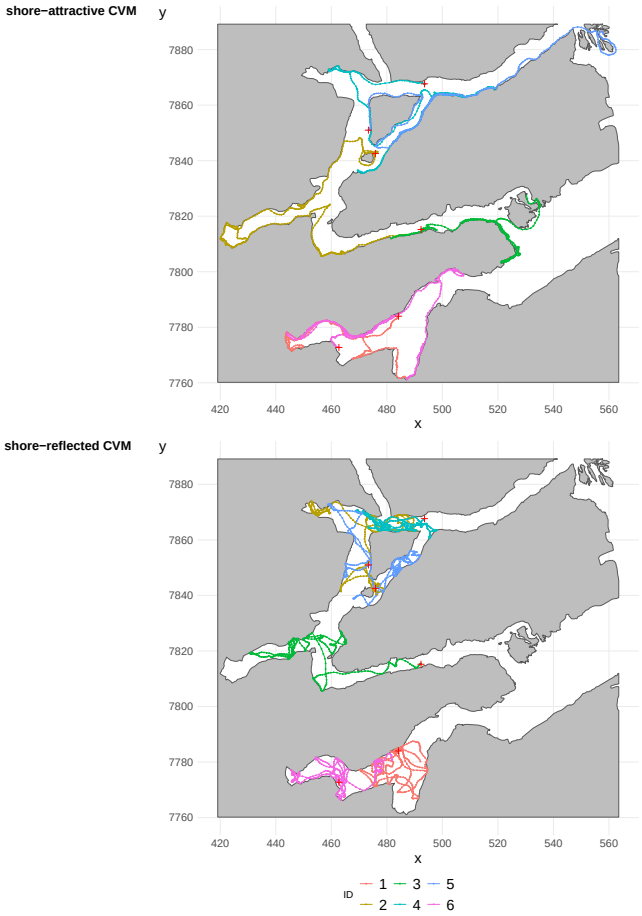


Figure 2: Trajectoires simulées sur trois jours pour le processus défini par l'EDS (1) avec la fonction (2) (en haut) et pour l'EDS réfléchie définie dans le papier de Hanks et al. (2017) (en bas). Les trajectoires sont simulées avec un pas de temps $\Delta = 10$ s, et sous-échantillonnées à 10 min.

$$y_{ij} = X_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2_{obs})$$

Les processus latents X_i et V_i sont solutions du système :

$$\begin{cases} dX_i(t) = V_i(t)dt \\ dV_i(t) = -\begin{pmatrix} \frac{1}{\tau_i} & -\omega_i(t) \\ \omega_i(t) & \frac{1}{\tau_i} \end{pmatrix} V_i(t)dt + \frac{2\nu_i}{\sqrt{\pi\tau_i}} dW(t) \end{cases}$$

où τ_i, ν_i sont des paramètres individuels :

$$\begin{pmatrix} \log(\tau_i) \\ \log(\nu_i) \end{pmatrix} = \begin{pmatrix} \log(\tau_0) \\ \log(\nu_0) \end{pmatrix} + \begin{pmatrix} b_{\tau,i} \\ b_{\nu,i} \end{pmatrix} \quad \text{avec} \quad \begin{pmatrix} b_{\tau,i} \\ b_{\nu,i} \end{pmatrix} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \begin{pmatrix} \sigma^2_\tau & 0 \\ 0 & \sigma^2_\nu \end{pmatrix}\right)$$

Les coefficients $\log(\tau_0)$ et $\log(\nu_0)$ sont des intercepts au niveau de la population, et $b_{\tau,i}$, $b_{\nu,i}$ sont les effets aléatoires individuels.

Enfin, comme discuté dans la Section 2, la fonction de vitesse angulaire de l'individu i, ω_i, est exprimée comme une fonction lisse de Θ et D^{shore} à l'aide de splines :

$$\omega_i(t) = \omega_0 + \sum_{k=1}^{L} \omega_k \psi_k(D_i^{shore}(t), \Theta_i(t)) \tag{3}$$

où L est le degré de liberté des splines. Les fonctions ψ_k dans l'équation (3) sont les fonctions de base bivariées construites à partir de fonctions de base univariées de splines cubiques. Nous renvoyons au livre de Wood (2017), section 4.1.8.

Le modèle de réponse est ensuite conçu pour évaluer une déviation par rapport au modèle de référence liée à l'exposition au navire. Nous ajoutons alors une dépendance à la covariable E^{ship} dans les paramètres τ_i et ν_i :

$$\begin{aligned}\log(\tau_i(t)) &= \log(\tau_0) + \alpha_\tau E_i^{ship}(t) + b_{\tau,i} \\ \log(\nu_i(t)) &= \log(\nu_0) + \alpha_\nu E_i^{ship}(t) + b_{\nu,i}\end{aligned}$$

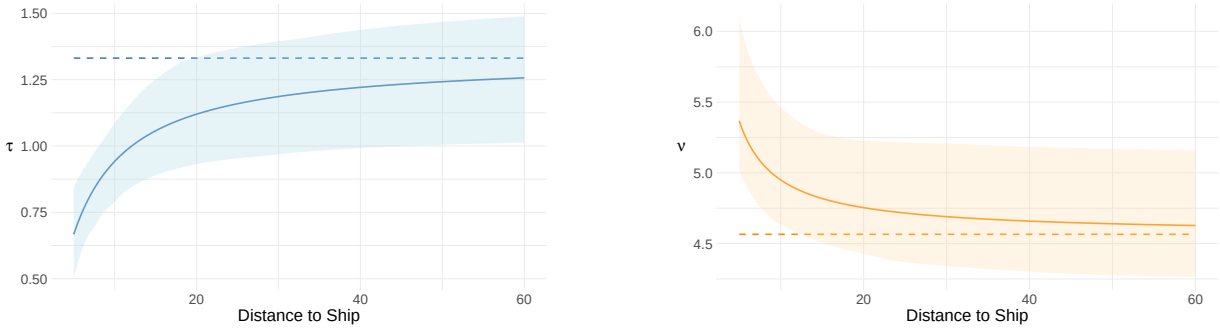
Les intercepts $\log(\tau_0)$, $\log(\nu_0)$ et la fonction lisse ω du modèle de référence interviennent comme termes fixes dans le modèle de réponse. Ainsi, seuls les coefficients α_τ et α_ν sont estimés à partir des données pendant l'exposition. Ces paramètres contrôlent la déviation de τ_i et ν_i par rapport à leurs valeurs de référence τ_0 et ν_0 en fonction de la distance au navire. Des coefficients α_τ et α_ν significativement différents de 0 indiquent un effet de l'exposition sonore sur le mouvement horizontal des narvals.

4 Estimation des paramètres

Pour l'estimation, nous faisons l'approximation suggérée dans l'article de Michelot (2021) que les paramètres de l'EDS sont constants sur chaque intervalle de temps $[t_{ij}, t_{i,j+1}]$, égaux à leur valeur au temps t_{ij}. Nous pouvons alors approximer la densité de transition des processus markoviens (X_i, V_i) pour obtenir les équations matricielles d'un modèle d'espace d'état linéaire gaussien. Cela nous permet d'utiliser le package `R smoothSDE` pour l'estimation des paramètres, basé sur l'algorithme de filtrage de Kalman. Cette méthode est déjà implémentée pour le cas $\omega = 0$ dans `smoothSDE`. Ici, nous l'étendons au modèle plus général (1) et l'utilisons pour l'estimation des modèles de référence et de réponse. Les résultats sont affichés dans le tableau 1. Les graphiques 3a et 3b montrent une diminution de la persistance du mouvement τ et une augmentation du paramètre de vitesse ν à mesure que le navire militaire se rapproche.

Paramètre	Estimation	IC
τ	1.34	$[1.06; 1.68]$
ν	4.57	$[4.08; 5.11]$
σ_τ	0.20	$[0.02; 0.79]$
σ_ν	0.10	$[0.03; 0.26]$
α_τ	-3.40	$[-5.31; -1.40]$
α_ν	0.91	$[0.18; 1.61]$

Table 1: Paramètres estimés avec intervalles de confiance à 95%



(a) (b)

Figure 3: Estimations de l'effet de la distance au navire sur les paramètres de population avec intervalles de confiance à 95%. Les pointillés sont les valeurs de référence estimées avant exposition. τ est en heures, ν en km/h et la distance au navire est en km.

Bibliographie

Heide-Jørgensen, Mads Peter et al. (2021), Behavioral Response Study on Seismic Airgun and Vessel Exposures in Narwhals *Frontiers in Marine Science*

Gurarie, Eliezer et al. (2017), Correlated velocity models as a fundamental unit of animal movement: synthesis and applications, *Movement ecology*

Hanks, Ephraim M. and Johnson, Devin S. and Hooten, Mevin B. (2017), Reflected Stochastic Differential Equation Models for Constrained Animal Movement, *Journal of Agricultural, Biological and Environmental Statistics*

Michelot, Théo and Glennie, Richard and Harris, Catriona and Thomas, Len (2021), Varying-Coefficient Stochastic Differential Equations with Applications in Ecology, *Journal of Agricultural, Biological and Environmental Statistics*

Modelling the phytoplankton community in a front: a Mediterranean Sea case study.

Théo Garcia [1], Laurina Oms [2], Xavier Milhaud [3], Andrea Doglioli [4], Monique Messié [5], Claire Lacour [6], Pierre Vandekerkhove [7], Gérald Grégori [8], Denys Pommeret [9]

[1] *Aix Marseille University CNRS, Centrale Marseille, I2M, Marseille, France, theo.garcia@univ-amu.fr*

[2] *Aix-Marseille Université, Université de Toulon, CNRS/INSU, IRD, Mediterranean Institute of Oceanography (MIO), UM 110, Campus universitaire de Luminy, Marseille, France, laurina.oms@mio.osupytheas.fr*

[3] *Aix Marseille University CNRS, Centrale Marseille, I2M, Marseille, France, xavier.milhaud@univ-amu.fr*

[4] *Aix-Marseille Université, Université de Toulon, CNRS/INSU, IRD, Mediterranean Institute of Oceanography (MIO), UM 110, Campus universitaire de Luminy, Marseille, France, andrea.doglioli@univ-amu.fr*

[5] *Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA, monique@mbari.org*

[6] *Laboratoire d'analyse et de mathématiques appliquées, Université Gustave Eiffel, 5 boulevard Descartes, 77454 Marne-la-Vallée Cedex 2, claire.lacour@univ-eiffel.fr*

[7] *Laboratoire d'analyse et de mathématiques appliquées, Université Gustave Eiffel, 5 boulevard Descartes, 77454 Marne-la-Vallée Cedex 2, pierre.vandekerkhove@univ-eiffel.fr*

[8] *Aix-Marseille Université, Université de Toulon, CNRS/INSU, IRD, Mediterranean Institute of Oceanography (MIO), UM 110, Campus universitaire de Luminy, Marseille, France, gerald.gregori@mio.osupytheas.fr*

[9] *Aix Marseille University CNRS, Centrale Marseille, I2M, Marseille, France, denys.pommeret@univ-amu.fr*

Résumé.

Les fronts sont des structures fines échelles (1-100 km) qui séparent différentes masses d'eau et créent des gradients de contenu biogéochimique. Les fronts sont souvent associés à un mélange vertical de la colonne d'eau, permettant la disponibilité de nutriments qui soutiennent la dynamique du phytoplancton. Cependant, la manière dont ces structures affectent la distribution du phytoplancton n'est pas bien comprise, en particulier dans les régions oligotrophes. Nous faisons l'hypothèse que la communauté phytoplanctonique observée dans le front est un mélange des communautés observées dans les masses d'eau adjacentes ainsi que d'une autre communauté.

Ici, nous nous intéressons à la composition de la communauté phytoplanctonique basée sur neuf types fonctionnels (PFT) observés par cytométrie de flux dans un front, dans l'ouest de la mer Méditerranée oligotrophe. Lors de la campagne PROTEVS MED, au sud des Baléares, des mesures à haute résolution nous ont permis de collecter des échantillons dans le front et les deux masses d'eau adjacentes, le long d'un fort gradient de salinité.

Notre objectif est de modéliser la communauté phytoplanctonique du front comme un mélange des communautés des masses d'eau adjacentes A et B, et d'une nouvelle communauté

C. Dans ce modèle, nous avons spécifié que les communautés des masses d'eau adjacentes et la nouvelle communauté peuvent provenir d'un mélange discret de distributions normales multivariées. Tout d'abord, nous avons estimé les paramètres et le nombre de composantes du mélange pour les communautés des masses d'eau adjacentes A et B à l'aide d'un algorithme espérance-maximisation (EM). À partir d'un ensemble de données plus important, nous avons estimé les paramètres d'un ensemble de communautés probables pour C. Ensuite, nous avons développé un modèle hiérarchique bayésien pour estimer le poids de chaque composantes du mélange discret. Enfin, le modèle hiérarchique bayésien a été exécuté une seconde fois en utilisant uniquement les composantes avec un plus grand poids pour la communauté C.

Une seule composante a suffi pour modéliser la communauté A (au nord du front), tandis que les communautés B (au sud du front) et C ont été modélisées à l'aide de deux composantes. La nouvelle communauté C explique une partie importante de la communauté frontale. Avec très peu d'observations dans la zone frontale (n=11), notre approche bayésienne a permis de mettre en évidence la distribution spatiale de la communauté phytoplanctonique autour du front. Notre résultat suggère que les conditions environnementales locales dans le front permettent l'émergence d'une nouvelle communauté. Ce travail est une première étape dans la compréhension des zones frontales dans une région oligotrophe, représentative de l'océan global.

Mots-clés. Modèle de Mélanges Gaussien, Modèle hiérarchique bayésien, Fines échelles, Océanographie.

Abstract.

Fronts are ubiquitous fine-scale physical features (1-100 km) that separate different water masses and create gradients of biogeochemical contents. Fronts are often associated with vertical mixing of the water column, allowing the availability of nutrients that support phytoplankton dynamics. However, how such structures affect phytoplankton distribution is not well understood, especially in oligotrophic regions. We hypothesize that the phytoplankton community observed in the frontal zone is a mixture of communities observed in the adjacent water masses, plus another community.

Here, we are interested in the community composition based on nine phytoplankton functional types (PFTs) observed by flow cytometry in a front, in the western oligotrophic Mediterranean Sea. During the PROTEVS MED cruise, south of the Balearic Islands, high-resolution measurements allowed us to collect samples in the front and the two adjacent water masses along a strong salinity gradient.

Our objective is to model the frontal phytoplankton community as a finite mixture of the adjacent water mass communities A and B, and a new community C. In this model, we specified that the communities in the adjacent water masses and the new community can arise from a discrete mixture of multivariate normal distributions. First, we estimated the parameters and number of components in the finite mixture for the adjacent water mass communities A and B using an Expectation Maximization algorithm. From a larger dataset, we estimated the parameters of a set of likely communities for C. Then, we developed a hierarchical Bayesian model to estimate the weight of each component of the discrete mixture. Finally, the hierarchical Bayesian model was run a second time using only the most significant

components for community C.

One component was sufficient to model community A (North to the front), while communities B (South to the front) and C were modeled with two components. The new community C explained a significant part of the frontal community. With very few observations in the frontal zone (n=11), our Bayesian approach highlighted the spatial distribution of the phytoplankton community around the front. Our result suggests that local environmental conditions in the front allow the emergence of a new community. This work is a first step in understanding frontal zones in an oligotrophic region, representative of the global ocean.

Keywords. Gaussian mixture models, Hierarchical Bayes model, Fine-scale structure, Oceanography.

1 Introduction

L'océan peut être vu comme une mosaïque de masses d'eau partageant des propriétés physiques, chimiques, biologiques communes séparées par des frontières. Les frontières entre différentes masses d'eaux sont appelées fronts. Dans les régions frontales, de forts gradients de contenu biogéochimique sont observés. Ces structures, dites fine-échelle, possèdent une emprise spatiale de l'ordre de 1 à 100 km et perdurent dans l'océan entre quelques jours et quelques semaines (Dickey, 2003). Bien que ponctuelles, ces zones jouent un rôle déterminant dans le fonctionnement des écosystèmes marins. En effet les fronts sont souvent associés à un mélange vertical de la colonne d'eau, permettant la disponibilité de nutriments qui supporte la dynamique des organismes photosynthétiques qui sont à la base de la chaîne alimentaire, le phytoplancton.

De part la spontanéité et la faible emprise spatiale de ces structures, peu d'informations existent sur la distribution des communautés phytoplanctoniques en zones de front, en particulier dans les zones oligotrophes et peu énergétiques. Dans ce travail nous faisons l'hypothèse que la communauté phytoplanctonique observée dans le front est issue d'un mélange des communautés observées dans les masses d'eau adjacentes ainsi que d'une autre communauté.

2 Méthodes

2.1 Site d'étude et données

Dans ce travail, nous analysons des données récoltées lors de la campagne PROTEVS MED (2018), en mai 2018, au sud des Baléares (Figure 1-A). L'objectif de cette campagne a été de rechercher des structures fine-échelles et de réaliser des mesures à haute résolution, afin d'obtenir des échantillons dans le front et les deux masses d'eau adjacentes.

Ici nous nous intéressons aux prélèvements phytoplanctoniques réalisés en surface sur l'hippodrome nord-sud, le long d'un fort gradient en salinité (Figure 1-B). Les masses d'eau

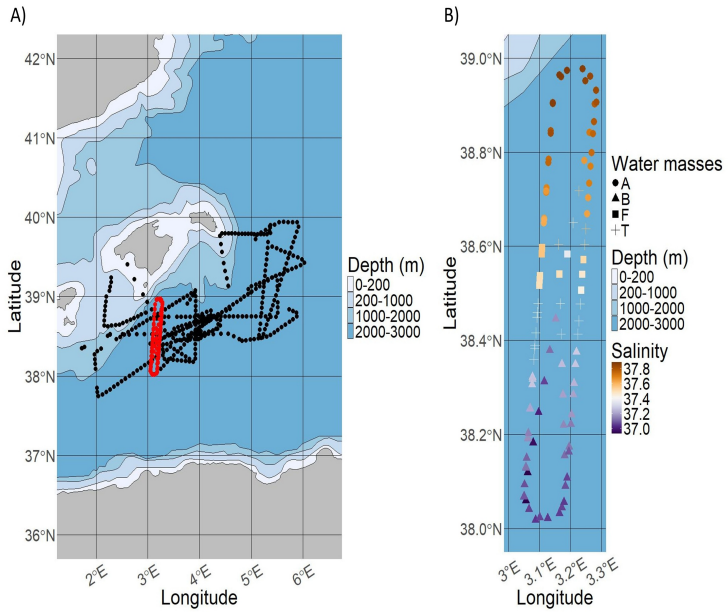


Figure 1: Carte du site d'étude. A) Les points (noirs et rouges) correspondent à chaque prélèvements d'eau pendant la campagne PROTEVS MED. Les points rouges correspondent à l'hippodrome Nord-Sud. B) Focus sur l'hippodrome Nord-Sud. Les points sont colorés selon leur salinité. La forme des points dépend de la masse d'eau à laquelle ils appartiennent.

ont été caractérisées selon leur salinité. Ici, A correspond à la masse d'eau du nord (salinité $\geq$ 37.6) et B à la masse d'eau du sud (salinité $\leq$ 37.3). Le front, F, a été défini par salinité (37.6 $\geq$ salinité $\geq$ 37.3) et par un critère géographique (38.6°$\geq$ Latitude $\geq$38.5°). Les stations avec une salinité comprise entre 37.3 et 37.6 mais avec une latitude $>$ 38.6 ° ou latitude $<$ 38.5° ont été considérées comme masse d'eau transitoire, T. En tout 30 échantillons ont été collectés pour la masse A, 44 pour la masse B, 11 pour le front F et 17 dans T.

Les échantillons ont été analysés par cytométrie en flux et ont permis la classifications des organismes observés en neufs types fonctionnels phytoplanctoniques (PFT) selons leur tailles et propriétés optiques (comme dans Tzortzis et al., 2021). Ces groupes sont *Synechococcus* (Syne), cryptophytes (Crypto), quatres groupes de Picoeucaryotes (Pico1, Pico2 and Pico3, PicoHFLR), deux groupes de Nanoplancton (SNano and RNano) et microphytoplancton (Micro). Pour chacun des échantillons, les biomasses des neuf PFT ont été estimées.

2.2 Modèle de mélange

Nous modélisons la répartition des biomasses de la communauté observée du front, F, comme un mélange discret des communautés A et B observées séparément, et d'une communauté inconnue, C, comme suit:

$$Y_F = \lambda_A Com^A + \lambda_B Com^B + \lambda_C Com^C, \quad (1)$$

avec Y_F une table de dimension n x m, comprennant $n = 11$ mesures de biomasses pour les $m = 9$ PFT. Les quantités inconnues λ_A, λ_B and λ_C sont les poids des trois composantes du mélange. Le cas où $\lambda_C \neq 0$, signifie l'existence d'une communauté nouvelle. En revanche, si $\lambda_C = 0$, le front est un mélange uniquement de A et B.

Une étude préliminaire a montré que la forte variabilité au sein des communautés A et B pouvait provenir d'un mélange. Pour cette raison, nous proposons que Com^A, Com^B et Com^C, sont elles mêmes issues d'un mélange de, respectivement, j, k et l, composantes de Gaussiennes multivarité:

$$Com^A = \alpha_{A1} MultiNorm(\mu_A^1, \Sigma_A^1) + ... + \alpha_{Aj} MultiNorm(\mu_A^j, \Sigma_A^j) \quad (2)$$

$$Com^B = \alpha_{B1} MultiNorm(\mu_B^1, \Sigma_B^1) + ... + \alpha_{Bk} MultiNorm(\mu_B^k, \Sigma_B^k) \quad (3)$$

$$Com^C = \alpha_{C1} MultiNorm(\mu_C^1, \Sigma_C^1) + ... + \alpha_{Cl} MultiNorm(\mu_C^l, \Sigma_C^l) \quad (4)$$

où α_{A1}, ...,α_{Aj}, α_{B1}, ..., α_{Bk}, α_{C1},..., α_{Cl} sont les poids des composantes du mélanges dans Com^A, Com^B and Com^C, respectivement. μ (vecteurs Gaussiens de taille $m = 9$) et Σ (matrices de covariance of dimension m x m) sont les paramètres des Gaussiennes multivariées.

Dans un premier temps les paramètres (μ et Σ) et le nombre de composantes dans Com^A, Com^B et Com^C ont été déterminés grâce à un algorithme EM. Pour Com^A et Com^B cette procédure a été réalisée sur les données observées respectivement en masse A et en masse B. Les données issues du reste de la campagnes, sans inclure l'hippodrome N-S, ont été utilisées pour estimer les l composantes possibles dans Com^C. Le nombre de composantes pour Com^A, Com^B et Com^C a été déterminé en utilisant le critère ICL (integrated complete-data likelihood).

Dans un deuxième temps un modèle hiérarchique bayésien a été utilisé pour estimer λ_A, λ_B, λ_C et α_{C1},..., α_{Cl}. Le modèle a été utilisé une première fois avec les l composantes dans Com^C. Puis, à partir des résultats de ce modèle, le modèle a été utilisé une deuxième fois en sélectionnant uniquement les composantes dans Com^C ayant le plus haut poids, α_C. L'échantillonnage des distributions *a posteriori* a été réalisé via l'algorithme Monte-Carlo Hamiltonien dans STAN. 4 chaînes de 11000 itérations ont été réalisées. Les 10000 premières itérations ont été considérés comme *burn-in*.

3 Résultats et discussion

Selon le critère ICL, une composante était suffisante pour modéliser Com^A, deux composantes pour Com^B ($B1$ et $B2$), et l=12 composantes probables ont été utilisées pour modéliser Com^C. Après avoir fait tourné le modèle une première fois, 2 composantes ont finalement été retenues pour modéliser Com^C (appelées $C8$ et $C10$). Sur le deuxième modèle, les valeurs des α de ces deux composantes $C8$ et $C10$ étaient respectivement 0.55 et 0.45 (Table 1). Parmis les trois communautés, Com^C présentait le plus grand poids dans le mélange (approx. 0.70), suivi de Com^A (approx. 0.22) et Com^B (approx. 0.06). Ces résultats montrent l'importance de Com^C dans le mélange validant ainsi l'hypothèse de départ.

Table 1: Résumé des distributions *a posteriori* des paramètres estimés par le modèle bayésien pour les quantiles 2.5, 25, 50, 75 et 97.5%.

	α_{C8}	α_{C10}	λ_A	λ_B	λ_C
2.5%	0.28	0.18	0.04	0	0.45
25%	0.45	0.35	0.13	0.03	0.63
50%	0.55	0.45	0.2	0.06	0.72
75%	0.65	0.55	0.28	0.12	0.8
97.5%	0.82	0.72	0.46	0.27	0.91

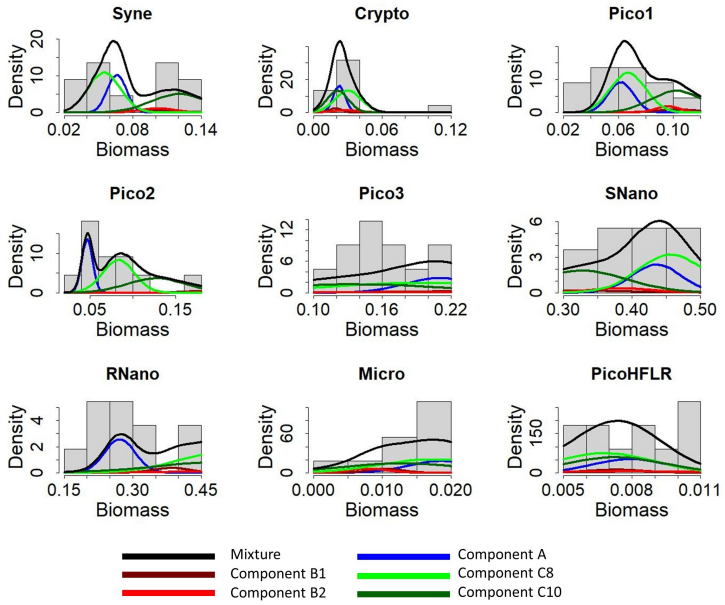


Figure 2: Ajustement des gaussiennes multivariées avec la biomasse observée pour chaque PFT dans la masse d'eau F. Les histogrammes montrent les densités de la biomasse des PFT, et les lignes correspondent aux courbes de densité des composantes gaussiennes multivariées estimées par EM et par le modèle bayésien. Notons que cinq composantes sont nécessaires pour modéliser la communauté F, chaque composante étant pondérée par ses valeurs λ et α. La ligne noire correspond au mélange (c'est-à-dire ici la somme) de ces cinq composantes.

Dans l'ensemble, le mélange des cinq composantes a permis de bien modéliser la distribution de la biomasse des PFT. Dans certains cas, les distributions de biomasse des PFT étaient multimodales (Syne, Pico2, RNano) ou asymétriques (Pico1, Micro), ce qui a justifié l'utilisation d'un mélange fini de gaussiennes multivariées. Cependant, pour certains PFT (Pico3 et RNano), une différence subsiste entre la densité estimée et la densité observée (Figure 2). Il pourrait s'agir de divergences entre les vraies communautés inconnues dans C et celle que nous avons proposée (en utilisant les données du reste de la campagne). Par exemple, on pourrait s'attendre à ce qu'avec une variance plus faible dans $C8$ et $C10$ pour Pico3, le mode autour de 0,14-0,16 aurait été mieux modélisé. Néanmoins, les nouvelles

composantes $C8$ et $C10$ de Com^C ont permis de mieux expliquer la communauté du front, que si elle n'avait été modélisée uniquement par Com^A et Com^B.

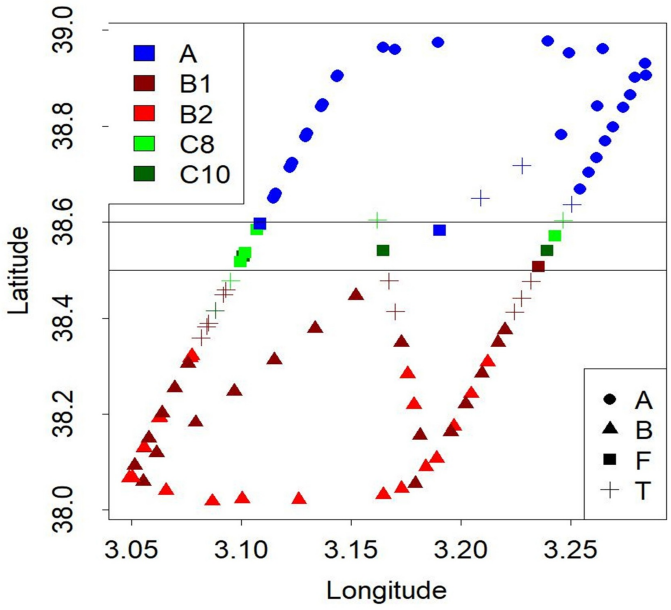


Figure 3: Distribution spatiale de la composante dominante dans les échantillons de l'hippodrome NS. Les points sont colorés en fonction de la composante gaussienne multivariée (A, $B1$, $B2$, $C8$ et $C10$) qui atteint la plus forte densité (en neuf dimensions) pour l'échantillon de biomasse observée par PFT. Les lignes horizontales noires correspondent à la zone frontale (i.e. latitude comprise entre 38,5°N et 38,6°N).

La figure 3 montre clairement la distribution spatiale de la composante dominante dans le mélange. La composante A est principalement dominante au nord de l'hippodrome, les composantes $B1$ et $B2$ sont principalement dominantes au sud de l'hippodrome, et les composantes $C8$ et $C10$ sont principalement dominantes dans le front. Seules quatre des observations dans les eaux de transition T, c'est-à-dire dans la gamme de salinité du front mais en dehors de la gamme latitudinale, sont dominées par $C8$ ou $C10$. Sur les onze observations dans le front F, deux sont dominées par A et une par $B1$. Ces résultats montrent qu'avec un faible nombre de données il possible de modéliser la communauté phytoplanctonique dans le front. De plus de part cette séparation spatiale des composantes gaussiennes multivariées, ces résultats suggèrent que des conditions environnementales permettent l'émergence d'une communauté nouvelle en zone de front.

Notre approche de modélisation sera appliquée à un ensemble de données plus large, notamment de la campagne BIOSWOT-MED (2023). Dans ces analyses ultérieures, des données environnementales seront incluses pour démêler les processus physico-biologiques qui façonnent la distribution du phytoplancton.

Remerciements

Ce travail a été financé par l'Institut des Mathématiques pour la Planète Terre qui soutient les collaborations entre mathématiciens et scientifiques en sciences de la vie et de la terre.

Bibliographie

Dickey, T. D., 2003. Emerging ocean observations for interdisciplinary data assimilation systems. *J.Marine Syst.* T. 40, p. 5-48. DOI : 10.1016/S0924- 7963(03)00011-3.

Doglioli, A.M., & Gregory, G. 2023. BioSWOT-Med cruise, *RV L'Atalante*. doi:10.17600/18002392.

Dumas, F., Garreau, P., Louazel, S., Correard, S., Fercoq, S., Le Menn, M., ... & Gregori, G. (2018). PROTEVS-MED field experiments: Very High Resolution Hydrographic Surveys in the Western Mediterranean Sea.

Tzortzis, R., Doglioli, A. M., Barrillon, S., Petrenko, A. A., d'Ovidio, F., Izard, L., ... & Gregori, G. (2021). Impact of moderately energetic fine-scale dynamics on the phytoplankton community structure in the western Mediterranean Sea. *Biogeosciences*, 18(24), 6455-6477.

Modélisation d'une épidémie végétale à transmission vectorielle sur graphes, application à l'épidémie de Flavescence dorée dans le vignoble de Bourgogne de 2013 à 2023.

Frédéric Fabre [1], Davide Martinetti[2] & Jérôme Roux [2]

[1] *UMR SAVE, INRAE, Bordeaux Sciences Agro, Villenave d'Ornon, France, frederic.fabre@inrae.fr*
[2] *BioSP (Biostatistique et Processus Spatiaux), INRAE, Avignon, France, davide.martinetti@inrae.fr*
[2] *BioSP (Biostatistique et Processus Spatiaux), INRAE, Avignon, France, jerome.roux@inrae.fr*

Résumé.

Dans ce travail nous explorons la dynamique de la Flavescence dorée (FD) dans le vignoble bourguignon (2013-2023) à travers la reconstruction des réseaux de contamination, étape cruciale pour comprendre la propagation spatio-temporelle des épidémies. La FD est une maladie vectorielle causée par un phytoplasme et transmise par la cicadelle Scaphoideus titanus, et elle constitue une menace majeure pour la viticulture européenne en provoquant des pertes de rendement et la mort des ceps contaminés.

Ce travail s'appuie sur le suivi de l'épidémie de FD au travers d'un jeu de données conséquent couvrant 100 000 parcelles sur dix ans, et sur la construction de deux réseaux de transmission complémentaires :

- Un réseau de proximité spatiale basé sur la capacité de dispersion naturelle du vecteur.
- Un réseau d'exploitation reflétant la dissémination passive via les activités humaines.

Dans ce travail nous montrons que le réseau de proximité seul ne peut pas expliquer la vitesse observée de propagation ni l'apparition de foyers secondaires distants mais qu'une fois couplé au réseau d'exploitation, favorisant une dissémination rapide sur de longues distances, ces deux réseaux se complètent.

Nous montrerons finalement qu'une adaptation d'un modèle simple d'épidémie de type SIS (Susceptible-Infecté-Susceptible) sur ces réseaux de transmission permet de simuler des épidémie réalistes une fois calibré. Cette étape de calibration du modèle par inférence des paramètres par maximisation de la log-vraisemblance visera également à quantifier la contribution relative des deux modes de transmission à l'épidémie empirique. Nous espérons que les résultats présentés contribueront à optimiser les stratégies de surveillance et de lutte contre la FD. Par la suite cette approche pourra évoluer avec l'ajout de nouveaux réseaux ou l'utilisation de graphes pondérés.

Mots-clés. Graphes de transmission, Inférence statistique, Modèle SIS, Flavescence dorée

Abstract. In this work, we explore the dynamics of Flavescence dorée (FD) in the Burgundy vineyards (2013-2023) through the reconstruction of contamination networks, a crucial step in understanding the spatio-temporal spread of epidemics. FD is a vector-borne disease caused by a phytoplasma and transmitted by the Scaphoideus titanus leafhopper, and poses a major threat to European viticulture by causing yield losses and the death of contaminated vines.

This work is based on monitoring the FD epidemic using a large dataset covering 100,000 plots over ten years, and on the construction of two complementary transmission networks:

- A spatial proximity network based on the vector's natural dispersal capacity.
- An exploitation network reflecting passive dissemination via human activities.

In this work, we show that the proximity network alone cannot explain the observed speed of spread or the appearance of distant secondary outbreak, but that when coupled with the exploitation network, which promotes rapid spread over long distances, these two networks complement each other.

Finally, we will show that adapting a simple SIS (Susceptible-Infected-Susceptible) epidemic model to these transmission networks enables realistic epidemics to be simulated once the model has been calibrated. This model calibration stage, using parameter inference by log-likelihood maximisation, will also aim to quantify the relative contribution of the two modes of transmission to the empirical epidemic. We hope that the results presented will help to optimise FD surveillance and control strategies. This approach may subsequently evolve with the addition of new networks or the use of weighted graphs.

Keywords. Transmission graphs, Statistical inference, SIS model, Flavescence dorée

La reconstruction des réseaux de contamination en épidémiologie constitue un enjeu majeur pour la compréhension des dynamiques spatio-temporelles des épidémies (Barbillon et al, 2020). L'identification de ces réseaux facilite en effet la sélection et le paramétrage des modèles épidémiques, permettant ainsi des analyses prospectives pour l'évaluation de différentes stratégies de surveillance et de gestion (Strona et al, 2017). Cette présentation porte sur une étude de la dynamique épidémique de la Flavescence dorée (FD) dans le vignoble bourguignon entre 2013 et 2023, selon trois axes principaux :

1 Présentation de l'épidémie et données disponibles

La Flavescence dorée, causée par un phytoplasme et transmise par la cicadelle Scaphoideus titanus (insecte vecteur), représente une menace significative pour la viticulture française et plus largement européenne. Cette maladie de quarantaine, soumise à une lutte obligatoire

en France, se caractérise par des symptômes jaunissement ou rougissement des feuilles selon les cépages, l'absence d'aoûtement des bois et la mortalité potentielle des ceps (Chuche et Thiéry, 2014). Ainsi, le dépérissement des plantes contaminées impacte significativement la production viticole par la réduction des rendements (jusqu'à 50 %) et la mortalité potentielle des ceps infectés (Chuche et Thiéry, 2014).

Pour cette étude les données analysées proviennent d'une surveillance épidémiologique systématique à l'initiative des viticulteurs, conduite sur une période décennale (de 2013 à 2023) et couvrant environ les 100 000 parcelles constituant le vignoble bourguignon. Cette base de données d'une ampleur conséquente est un atout considérable à l'analyse des processus épidémique de la FD mais présente néanmoins un biais potentiel lié à la présence concomitante du Bois noir, une maladie présentant une symptomatologie similaire, induisant un risque de faux négatifs qui devra être intégré dans l'analyse statistique.

2 Construction des réseaux de transmission

La modélisation des voies de transmission repose sur deux réseaux complémentaires :

- Réseau de proximité spatiale :

  Ce réseau est modélisé par un graphe où les nœuds représentent les parcelles et une arête existe entre deux parcelles si leur distance est inférieure à un certain seuil. Le seuil de distance repose ici sur la capacité de dispersion de l'insecte vecteur, ici fixé à quelques dizaine de mètres ($\approx$ 30m)(Chuche et Thiéry, 2014). Ainsi ce réseau modélise les chemins que peut activement emprunter l'insecte vecteur.

- Réseau d'exploitation viticole :

  Ce réseau est modélisé par un graphe où les nœuds représentent les parcelles et une arête existe entre deux parcelles si elles sont rattachées à la même exploitation viticole. Ainsi ce réseau modélise les chemins que peut emprunter l'insecte vecteur selon une dissémination passive via les activité humaines.

Une analyse comparative de la topologie de ces réseaux (binaires non orientés)révèle que le réseau de proximité seul ne peut expliquer la vitesse de propagation observée ni l'apparition de foyers épidémique secondaires distants. Le réseau d'exploitation présente en revanche une topologie favorisant une dissémination rapide sur de longues distances et vient compléter la dissémination "locale".

3 Modélisation épidémiologique et inférence statistique

Pour expliquer l'épidémie observée nous proposons le développement d'un modèle épidémique de type SIS (Susceptible-Infecté-Susceptible)(Ahn et al, 2013) caractérisée par 4 paramètres:

- P_p : probabilité de transmission par voie vectorielle.

- P_e : probabilité de transmission via le réseau d'exploitation.
- α : coefficient d'efficacité des traitements insecticides (modulant P_p et P_e).
- P_r : probabilité d'assainissement des parcelles détectées infectées.

Et couplé à un processus d'observation avec :

- P_d : probabilité de détection d'une parcelle infectée, permettant la modélisation des faux négatifs.

L'inférence des paramètres est réalisée par maximisation de la log-vraisemblance (Barbillon et al, 2020) sur les données empiriques de l'épidémie bourguignonne. Cette approche vise ainsi à quantifier la contribution relative des deux modes de transmission dans la dynamique épidémique observée. Cette approche pourra être amenée à évoluer, notamment par l'ajout de nouveau réseau de transmission (par exemple routier), ou par l'utilisation de graphes pondérés autorisant une plus grande flexibilité dans l'existence d'un chemin de transmission.

Cette étude propose ainsi un cadre méthodologique original pour l'analyse des épidémies à transmission vectorielle en contexte agricole, combinant modélisation mécaniste et inférence statistique. A terme, les résultats obtenus pourront contribuer à l'optimisation des stratégies de surveillance et de lutte contre la Flavescence dorée notamment par l'identification de parcelles "sentinelles" (Strona et al, 2017) ou la préconisation de traitement insecticide.

Bibliographie

Ahn, H. J., et Hassibi, B. (2013, December). Global dynamics of epidemic spread over complex networks. In 52nd IEEE Conference on Decision and Control (pp. 4579-4585). IEEE.

Barbillon, P., Schwaller, L., Robin, S., Flachs, A., et Stone, G. D. (2020). Epidemiologic network inference. Statistics and Computing, 30(1), 61-75.

Chuche, J., et Thiéry, D. (2014). Biology and ecology of the Flavescence dorée vector Scaphoideus titanus: a review. Agronomy for sustainable development, 34, 381-403.

Strona, G., Carstens, C. J., et Beck, P. S. (2017). Network analysis reveals why Xylella fastidiosa will persist in Europe. Scientific Reports, 7(1), 71.

Amélioration de cartes de concentrations en NO_2 dans l'air, par combinaison de mesures de qualités différentes

Camille Coron [1] & Emma Thulliez [2]

[1] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France. e-mail : camille.coron@inrae.fr*
[2] *INSA Rouen Normandie, Normandie Univ, LMI UR 3226, F-76000 Rouen, France. e-mail : emma.thulliez@insa-rouen.fr*

Résumé. La qualité de l'air urbain est un enjeu majeur aujourd'hui. Les concentrations en polluant, comme le NO_2, doivent être surveillées afin de vérifier qu'elles ne dépassent pas les seuils de dangerosité. Deux techniques récentes aident à cartographier les concentrations en polluants à petite échelle. D'abord, les modèles physico-chimiques déterministes permettent de considérer le réseau de rues et calculent des estimations de la concentrations sur un maillage, donnant une carte. D'autre part, l'avènement de nouvelles technologies à bas-coût permet aux organismes de surveillance de densifier les réseaux de mesures. Ces appareils sont cependant moins fiables que les appareils de référence, et doivent être corrigés. Nous proposons une approche nouvelle pour améliorer les cartes issues des modèles déterministes en combinant les mesures de plusieurs réseaux de capteurs. Pour cela, nous modélisons le biais des modèles déterministes et l'estimons avec une méthode MCMC. Cela nous permet de tenir compte de nouvelles variables et de proposer une carte corrigée dans un contexte de prévision.

Mots-clés. Qualité de l'air, Cartographie, Estimation de biais, fusion de données, MCMC, Calibrage de capteur, Dioxyde d'azote

Abstract. Urban air quality is a major issue today. Pollutant concentrations, such as NO_2, must be monitored to ensure that they do not exceed dangerous thresholds. Two recent techniques help map pollutant concentrations on a small scale. First, deterministic physico-chemical models take into account the street network and calculate concentration estimates on a grid, providing a map. On the other hand, the advent of new low-cost technologies allows monitoring organizations to densify measurement networks. However, these devices are less reliable than reference devices and need to be corrected. We propose a new approach to improve maps generated using deterministic models by combining measurements from multiple sensor networks. To do this, we model the bias of deterministic models and estimate it using an MCMC method. This allows us to account for new variables and provide a corrected map in a forecasting context.

Keywords. Air quality, Mapping, Bias estimation, data fusion, MCMC, Sensor calibration, Nitrogen dioxide

1 Introduction et motivations

Ces dernières années, la qualité de l'air est apparue comme un enjeu majeur pour notre société. D'une part, l'OMS alerte sur l'impact de la pollution sur la santé et d'autre part, les institutions européennes comme françaises établissent et mettent à jour des réglementations sur les concentrations en polluants. Il est donc nécessaire d'établir d'avoir des estimations fiables de la concentration en polluants. En zone urbaine, cela concerne notamment le dioxyde d'azote (NO_2) et les particules fines.

Afin d'avoir une idée de la concentration en NO_2 à petite échelle, les organismes de surveillance utilisent des cartes de concentration, issues de modèles physico-chimiques déterministes. Ces modèles sont améliorés régulièrement mais sont génériques, et ne s'adaptent pas nécessairement bien à un lieu particulier. En effet, ils ne prennent pas en compte certaines spécificités géographiques comme l'altitude, ou temporelles comme les travaux urbains.

Par ailleurs, les grandes villes disposent également d'instruments de mesures pour surveiller la qualité de l'air. Les plus fiables, les stations de références, sont très coûteuse et donc peu nombreuses (entre 4 et 10 par agglomération). Leur faible densité spatiale est de plus en plus compensée par la mise en service de micro-capteurs, mais dont les mesures sont moins fiables, biaisées et fortement bruitées.

Notre approche consiste à combiner les données de mesures de ces différents instruments, en vue d'estimer les erreurs des sorties de modèles physico-chimiques, et de fournir ainsi des cartes corrigées avec un modèle spécifique au lieu considéré. Contrairement à un travail précédent [1], l'approche propose une correction globale et non différenciée par zones.

Le plan de cet article est le suivant : nous présentons les données, puis le modèle étudié, sa méthode d'estimation et enfin quelques résultats.

2 Données

Sorties de modèles physico-chimiques déterministes Nous utilisons des cartes issues du modèle SIRANE [2, 3, 4], qui donnent des concentrations horaires de NO_2 à Rouen, avec une précision de 10m×10m. Un exemple de carte est donné en Figure 2.

Mesures de concentrations Les concentrations dans l'agglomération de Rouen sont mesurées à travers 4 stations de référence (2 de trafic et 2 de fond). Elles sont complétées par 10 micro-capteurs AirSensEUR [5] installés sur des feux, dont les mesures en NO_2 sont exprimées en μV. Les micro-capteurs mesurent également le NO, l'O_3, le CO, la pression atmosphérique, l'humidité relative et la température. La Figure 1 présente la répartition du réseau dans la ville (une station est exclue car trop excentrée).

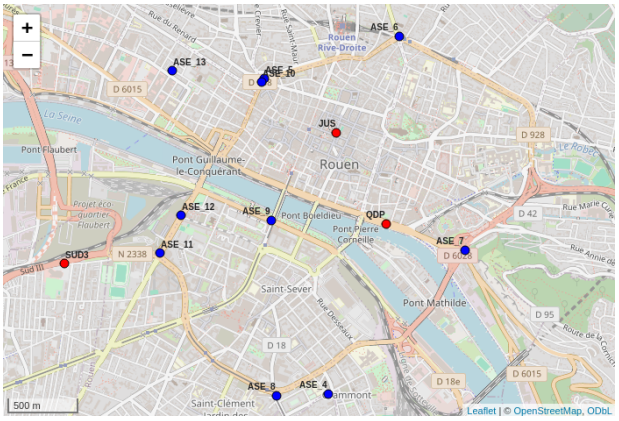


FIGURE 1 – Carte des emplacements des micro-capteurs et stations fixes à Rouen.

Mesures de covariables spatio-temporelles Pour caractériser la ville de Rouen, nous considérons des variables spatiales, qui sont la présence des espaces verts et de trafic [6] et l'altitude [7]. Nous incluons également des variables temporelles utilisées en entrée de SIRANE (température, frottement du vent).

3 Modèle

Cartes de concentrations Notons $s \in \mathbb{R}^2$ une coordonnées, $t \in \mathbb{R}^+$ un instant. Le principe de notre approche est de supposer que, les modèles déterministes physico-chimiques négligeant certains aspects spatio-temporels propres à chaque ville, les cartes qui en sont issues présentent des biais dont nous estimerons les paramètres. En notant $M(s,t)$ le modèle déterministe, $C(s,t)$ la concentration réelle en polluant, et $B(s,t)$ le biais, nous avons

$$M(s,t) = C(s,t) + B(s,t) \tag{1}$$

Plus spécifiquement, nous supposons que le biais B que nous allons chercher à estimer est de la forme :

$$B(s,t) = a_0 + \theta_T X_T(t) + C(s,t)\,(a_c + \zeta_S X_S(s) + \zeta_T X_T(t)) \tag{2}$$

où X_T est le vecteur des l observations temporelles et X_S le vecteur des k observations spatiales. Les coefficients associés à ces informations sont notés $\zeta_T, \theta_T \in \mathbb{R}^l$ et $\zeta_S \in \mathbb{R}^k$. Enfin, a_0 et a_c sont des scalaires, tels que $a_c \neq 0$.

Mesures Notons i l'indice d'une station fixe et $s_{0,i}$ sa position. La mesure $Z_{0,i}(t)$ réalisée par cette station fixe au temps t est reliée à la concentration réelle $C(s_{0,i},t)$ par l'équation :

$$Z_{0,i}(t) = C(s_{0,i},t) + \varepsilon(s_{0,i},t) \quad \text{où} \quad \varepsilon(s_{0,i},t) \sim \mathcal{N}(0,\sigma_0) \tag{3}$$

Pour le micro-capteur j, on suppose que la transformation entre mesure en μV ($Z_{1,j}(t)$) et concentration en $\mu g{\cdot}m^{-3}$ est linéaire et dépendante des autres polluants et caractéristiques météo (informations du vecteur $Y_j(t)$).

$$Z_{1,j}(t) = \frac{1}{\alpha_j}\left(C(s_{1,j},t) - \beta_j - \gamma_j Y_j(t) + \varepsilon_j(s_{1,j},t)\right) \quad \text{où } \alpha_j < 0 \text{ et } \varepsilon(s,t) \sim \mathcal{N}(0,\sigma_j) \qquad (4)$$

4 Procédure d'estimation

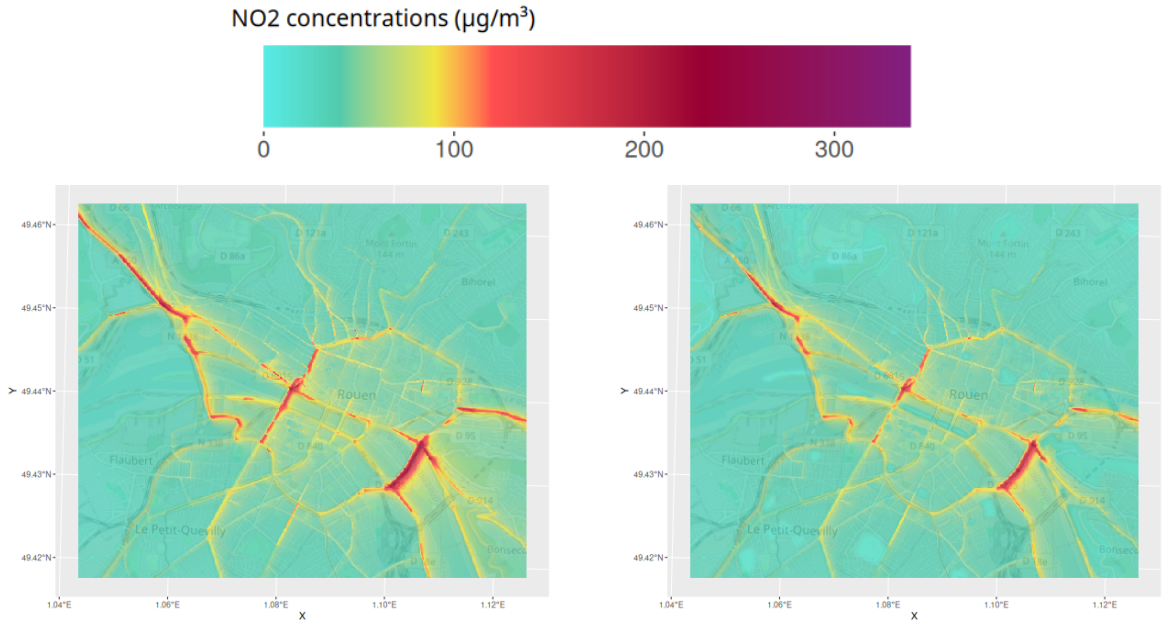
Les modélisations du biais et des mesures sont toutes deux dépendantes de la concentration réelle. Ainsi, la procédure d'estimation tient compte de cette imbrication. On propose d'estimer les paramètres du modèle avec une approche Markov-Chains Monte-Carlo (MCMC). On se donne des lois *a priori* sur les paramètres, puis on applique les étapes suivantes :

1. tirer des paramètres à partir de la loi *a priori*.
2. générer des simulations de la concentration à partir des équations (1) et (2).
3. établir des simulations des mesures à partir des équations (3) et (4).
4. comparer mesures observées et mesures simulées.

On peut ensuite établir des lois *a posteriori* des paramètres pour obtenir les paramètres estimés.

5 Résultats

La procédure nous permet de construire des cartes corrigées de la concentration en NO_2. Par exemple, dans un cas de sur-estimation par SIRANE de la concentration en NO_2, on obtient la correction proposée en Figure 2. La carte corrigée présente des concentrations estimées plus faibles. Notre approche permet aussi d'estimer les paramètres de transformation des micro-capteurs, donnés dans l'Équation (4), et par exemple d'étudier leur évolution au cours du temps, ce qui est une question importante pour les fabricants.



a) Carte avant correction b) Carte après correction

FIGURE 2 – Exemple de correction d'une carte SIRANE (07/12/2022, 8h)

Enfin, pour valider la méthode, nous simulons la situation en un point du domaine en retirant une des stations fixes de nos observations. Nous estimons les paramètres et regardons ensuite la qualité de la correction en cette station.

Références

[1] Benjamin Auder, Camille Coron, Jean-Michel Poggi, and Emma Thulliez. Debiasing physico-chemical models in air quality monitoring by combining different pollutant concentration measures. working paper or preprint, August 2024.

[2] Lionel Soulhac, Pietro Salizzoni, F.-X. Cierco, and Richard Perkins. The model sirane for atmospheric urban pollutant dispersion ; part i, presentation of the model. *Atmospheric Environment*, 45(39) :7379–7395, 2011.

[3] Lionel Soulhac, Pietro Salizzoni, P. Mejean, D. Didier, and I. Rios. The model sirane for atmospheric urban pollutant dispersion ; part ii, validation of the model on a real case study. *Atmospheric Environment*, 49 :320–337, 2012.

[4] Lionel Soulhac, C.V. Nguyen, Perrine Volta, and Pietro Salizzoni. The model sirane for atmospheric urban pollutant dispersion. part iii : Validation against no2 yearly concentration measurements in a large urban agglomeration. *Atmospheric Environment*, 167 :377–388, 2017.

[5] M Gerboles, M Signorini, and L Spinelle. Airsenseur – an open data/software /hardware multi-sensor platform for air quality monitoring. part a, sensor shield. Technical report, Joint Research Centre and Institute for Environment and Sustainability and Kotsev, 2015.

[6] OpenStreetMap contributors. Planet dump retrieved from https ://planet.osm.org . `https://www.openstreetmap.org`, 2017.

[7] Institut National de l'Information Géographique et Forestière (IGN). BD ALTI 2.0. `https://geoservices.ign.fr/bdalti`, 2024.

Environnement et statistique – Données de pluies

Spatio-temporal generation of precipitations using a spatially correlated Bernoulli and hidden Markov model

Caroline Cognot [1&2] & Liliane Bel [2] & David Métivier [3] & Sylvie Parey [1]

[1] *EDF R&D, 91120, Palaiseau, France - caroline.cognot@edf.fr*
[2] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France*
[3] *UMR MISTEA, 34000, Montpellier, France*

Résumé. Ce travail propose un modèle spatio-temporel pour l'occurence et l'intensité des précipitations, étendant le modèle de Markov caché multivarié utilisé dans la génération de données météorologiques. Dans les travaux précédents, l'occurrence de pluie à chaque site est modélisée indépendamment, conditionnée par un état caché commun représentant un type de temps. Dans notre approche, l'occurrence spatiale de la pluie est modélisée à l'aide d'un champ gaussien avec une structure de covariance de Matérn, seuillée en une variable de Bernoulli en fonction de la probabilité locale de pluie. Ce modèle spatial est intégré dans un modèle de Markov caché permettant d'identifier des états ou type de temps interprétables, tels que la probabilité de pluie ou l'extension spatiale. Nous intégrons des paramétrisations périodiques pour capturer les schémas saisonniers de manière efficace, plutôt que d'ajuster des modèles séparés par mois ou par saison. La quantité de pluie est modélisée conditionnellement à l'état météorologique à l'aide de distributions adaptées. Les paramètres du modèle sont estimés à l'aide de la vraisemblance composite par paires et de l'algorithme EM. Les simulations issues du modèle ajusté sont comparées aux données observées à l'aide de plusieurs métriques, comme le ratio d'occurrence de pluie (ROR). Le modèle utilise des données de 37 stations météorologiques en France, et 51 ans de données quotidiennes de précipitations. Les premiers résultats indiquent une portée spatiale estimée plus importante en hiver qu'en été, ce qui est conforme aux connaissances météorologiques attribuant les précipitations hivernales à des phénomènes à grande échelle et celles estivales à des processus à plus petite échelle. Notre modèle permet de générer des précipitations à haute résolution, avec des schémas météorologiques réalistes reproduisant des phénomènes clés telles que les épisodes de sécheresse.

Mots-clés. Générateurs de temps, Chaînes de Markov cachées, Spatio-temporel, Algorithme EM

Abstract. This study proposes a spatio-temporal model for precipitation across multiple stations, extending the multivariate hidden Markov model used in weather generation to a spatial framework for both occurrence and intensity. In previous works, rain occurrence at each site is modeled independently, conditioned on a shared hidden state representing a weather type. In our approach, spatial rain occurrence is modeled using a Gaussian field with a Matérn covariance structure, thresholded at each location into a Bernoulli variable based on the local rain probability. We incorporate this spatial model in the hidden Markov

framework, with the aim of identifying interpretable states, such as those characterized by lower rain probability or higher spatial correlation. Additionally, we incorporate periodic parameterizations to capture seasonal patterns efficiently, rather than fitting separate models for each month. The rain quantity is modeled conditionally to the weather state using appropriate distributions. Model parameters are estimated using composite pairwise likelihood and the EM algorithm. Simulations from the fitted model are compared to observed data using several metrics such as the rain occurrence ratio (ROR), which represents the daily proportion of stations experiencing rainfall. The model is applied to 37 weather stations in France, using 51 years of daily precipitation data. Initial results from the spatial model indicate a higher estimated spatial range in winter compared to summer. This aligns with meteorological knowledge that attributes winter rainfall to large-scale phenomena and summer rainfall to smaller-scale processes. Our model allows the generation of high-resolution rainfall with realistic weather patterns reproducing key events such as droughts.

Keywords. Weather Generator, Hidden Markov Model, Spatio-temporal, EM algorithm

1 Introduction

Weather generators consider the observations of a meteorological variable to be one of the infinite possible trajectories of a stochastic model. Their main objective is to allow the simulation of many plausible sequences of a meteorological variable, that share common statistical properties with the observations. They provide a complementary sampling of climate variables to physical models. Since the first weather generator of Richardson [1981] many others have been proposed using various statistical tools, for univariate and multivariate series. A common modeling choice is to separate dry and wet days, leading to the class of weather type models. Among these, hidden Markov models (HMMs) have long been popular for studying meteorological variables. As highlighted in Touron [2019], their key advantage is the conditional independence assumption: observed variables (temperature, precipitations, wind speed) are independent given a hidden state representing weather types. This assumption is also leveraged in Gobet et al. [2024] for a multisite model where the rain occurence are spatially independant conditionnaly to a hidden state.

Meanwhile, methods using latent or transformed Gaussian fields to model spatially correlated data have gained attraction. For instance, Evin et al. [2019] employ a probability transformation function to map a Gaussian field to observed variables. Combining this approach with the HMM framework, Ailliot et al. [2009] proposed a model with spatially correlated emissions, where rain occurrence follows a transformed Gaussian field. However, their approach is designed for a small number of stations and treats each month separately due to the computational complexity of the likelihood estimation.

In this work, we extend the HMM model of Gobet et al. [2024] by incorporating spatially correlated emissions. We apply the methodology to a network of 37 stations in France and provide interpretability in the estimated hidden states.

2 Methods

2.1 Model definition

We propose a Spatial Hidden Markov Model, presented in figure 1. At each time step, a Markov variable denotes the weather type, conditionally to which the observed variables are generated using a spatial gaussian field.

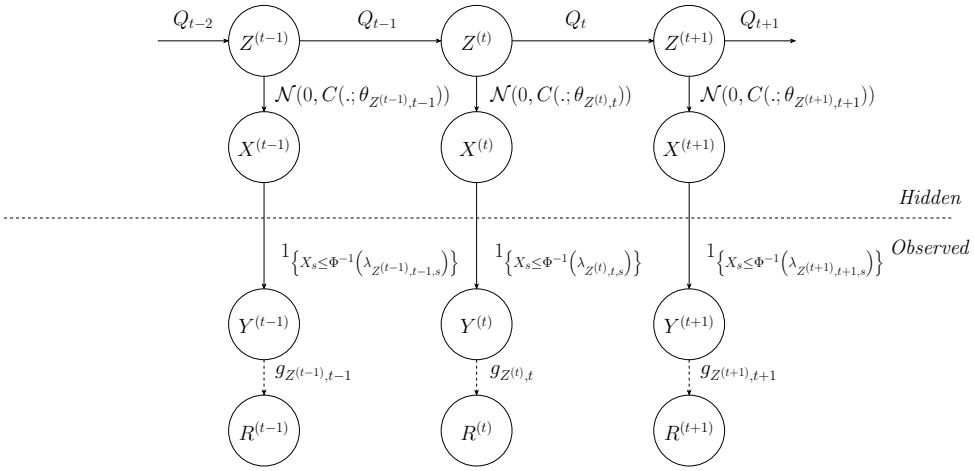


Figure 1: Hidden Markov Model linking hidden states (= weather types) Z to multisite correlated rain occurrence Y and rain intensity R.

Rain Occurrence Let $t \in \{1, \ldots, n_T\}$ be a time step, and let $s_1, \ldots, s_{d_S}$ denote d_S spatial points, with d_{ij} representing the distance between sites i and j.

The multisite correlated rain occurrence at time t, denoted by $\left(Y_i^{(t)}\right)_{i \in \{1,\ldots,d_S\}}$, is based on an unobserved Gaussian vector with a covariance function given by

$$C(h) = \mathcal{M}(h; \rho_{Z^{(t)},t}, \nu_{Z^{(t)},t}), \tag{1}$$

where $\mathcal{M}$ is the Matérn function, $\rho_{Z^{(t)},t}$ is the spatial range parameter, and $\nu_{Z^{(t)},t}$ is the regularity parameter associated with the hidden state at time t.

The rain occurrence is obtained via the transformation

$$\forall i \in \{1, \ldots, d_S\}, \quad Y_i^{(t)} = \begin{cases} 1, & \text{if } X_i^{(t)} \leq \Phi^{-1}(\lambda_{Z^{(t)},t,i}), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Here, Φ denotes the cumulative distribution function (CDF) of the standard normal distribution. $\lambda_{Z^{(t)},t,i}$ represents the rain probability in the hidden state at time t.

Rain Intensity Conditionnaly to the rain occurrence, rain intensity can be incorporated into the model using several approaches:

- Compute the Maximum A Posteriori (MAP) sequence of hidden states, then fit a rain intensity model for each class. Common choices for marginal distributions include Gamma and double-exponential models.
- Modify the model as in Ailliot et al. [2009]: instead of obtaining only 0 or 1 after transformation, allow for 0 or a positive value by applying a transformation to the Gaussian vector.
- Integrate the rain intensity model directly within the HMM framework and adjust the Expectation-Maximization (EM) algorithm accordingly.

2.2 Parameter estimation : rain occurrence

Likelihood of the model For n_T times steps, d_S locations, and a set of parameters θ representing all the parameters of the model, the likelihood of the observed rain occurrences is

$$L\left[\left(y^{(1)},\ldots,y^{(n)}\right);\theta\right] = \sum_{(z^{(1)},\ldots,z^{(n)})} \left(\Pi_{t=1}^{n} L(y^{(t)}|z^{(t)},\theta)\right)\left(\pi(z^{(1)})\Pi_{t=1}^{n-1} Q_t(z^{(t)},z^{(t+1)})\right) \tag{3}$$

where Q_t is the transition matrix at time t, π is the distribution of the initial states, and the likelihood of the vector $y^{(t)} = (y_1^{(t)},...,y_{d_S}^{(t)})$ given the hidden state $z^{(t)}$ is

$$L(y^{(t)}|z^{(t)},\theta) = \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} f_\Sigma((u_1,...,u_d))d(u_1,...,u_d) \tag{4}$$

with $a_i = -\infty$ if $Y_i = 1, \Phi^{-1}(\lambda_{Z^{(t)},t,i})$ else, $b_i = \infty$ if $Y_i = 0, \Phi^{-1}(\lambda_{Z^{(t)},t,i})$ else, and f_Σ is the density of the Gaussian Multivariate density, with covariance function $\mathcal{M}(.;\rho_{Z^{(t)},t},\nu_{Z^{(t)},t})$.

The EM algorithm The EM algorithm [Dempster et al., 1977] can be used to obtain the parameters of the model. The M-step involves maximizing a weighted sum of $\log L((y_1^{(t)},...,y_d^{(t)});\theta_{t,k})$ for all time steps t and all possible states k.

The $\lambda_{Z^{(t)},t,i}$ are taken as the empirical probabilities, independently from each other, inside each M step.

The likelihood is too complicated to compute. Instead in the Maximisation step of the algorithm, we use the pairwise likelihood : $\log L(y_1^{(t)},...,y_d^{(t)});\theta_{t,k}$ is replaced with

$$\log L^{WC}((y_1^{(t)},...,y_d^{(t)});\theta_{t,k}) = \sum_{i,j} w_{ij} \log L((y_i^{(t)},y_j^{(t)});\theta_{t,k}) \tag{5}$$

This procedure is known as Pairwise-EM or PEM in Varin et al. [2005].

The weights w_{ij} are chosen to reduce the number of terms and remove distant pairs from the estimation. As Bourotte [2016] we choose $w_{ij} = 0$ if $d_{ij} > t_d$, 1 else.

2.3 Validation of the model

To assess the validity of the weather generator, several indicators are computed for many simulations and compared to those obtained from the observations.

Local indicators For each location i, the rain intensity distribution, the dry spell length and the rain probability for each month can be obtained easily.

Spatial and spatio-temporal indicators Useful indicators are the correlation in rain intensity between pairs of stations at various time lags, $\mathrm{Cor}(R(s_i, \cdot + \Delta t), R(s_j, \cdot))$.

For the spatial rain occurence, we define the Rain Occurrence Ratio (ROR) as

$$ROR(t) = \frac{\sum_{s \in \mathcal{S}} Y_s^{(t)}}{|\mathcal{S}|} \tag{6}$$

This series shows the proportion of stations that experience rainfall at time t. In turn, we evaluate the model by computing the distribution of the ROR and its lagged autocorrelation.

In the same manner, we define the Rain Exceedance Ratio (RER) as

$$RER_\alpha(t) = \frac{\sum_{s \in \mathcal{S}} 1_{\{R_s^{(t)} > q_s^\alpha\}}}{|\mathcal{S}|} \tag{7}$$

Where q_s^α is the quantile of order α at station s.

3 Results : application to French station data

Data The ECA&D dataset (Klein Tank [2002]) is a high-quality European dataset of observed daily station data, covering periods that range from centuries for some stations to decades for others. We use a set of 37 weather stations in France, with high quality data available from 1973 to 2023 (51 years), see figure 2 for the locations.

Spatial rain occurrence model First, the model without temporal evolution states is fitted on the data, separating the data by months. Only $d_S + 2$ parameters for each month need to be estimated. These will serve as the base for the estimation of the full model. Initial estimations showed that fixing the smoothness parameter ν at $1/2$ was a good choice, so the only spatial parameter to estimate is the range parameter ρ.

Two versions of the model were fitted to the data. The first one only has one class, the second one has four classes. For the model with a single class, the spatial range is estimated to be higher in winter than in summer, which is coherent with meteorological knowledge: in winter, rain is caused by larger-scale phenomena than in summer. For the model with four classes, the estimated ranges are lower, indicating that some of the spatial correlation is supported by the hidden state.

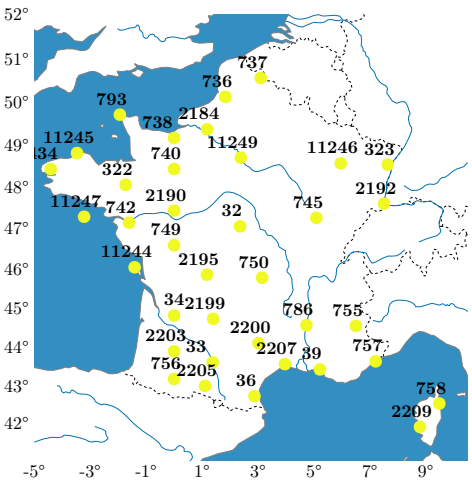


Figure 2: Map and location of the 37 stations.

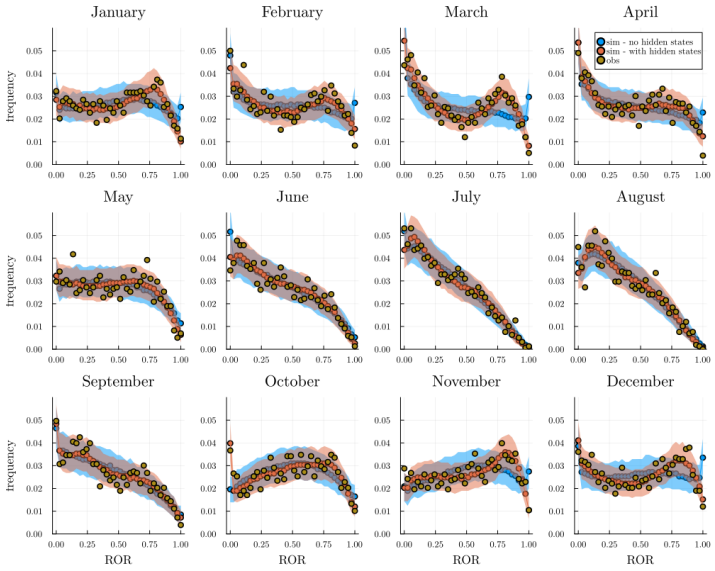


Figure 3: Observed *ROR* distribution, for each month, compared with 500 simulations from a model with only one class and a model with 4 classes.

The rain probability for each location is well represented compared to the empirical probabilities.

Figure 3 shows the distribution of the Rain Occurrence Rate (ROR) for each month, for the model without hidden states and the model with four states. The model with hidden states accurately represents the bimodal distribution in winter, while the model without hidden states does not. This bimodal property is a key information for French precipitation data : in the winter, the rain events are associated with large-scale atmospheric circulations, so there are two main observed situations : rain almost everywhere, and rain almost nowhere.

4 Conclusion and perspectives

In this work, we have developed a spatio-temporal weather generator for precipitation and proposed validation criteria. By kriging the local parameter, our model can be used to generate precipitation data for any set of spatial points.

Currently, the model is implemented only for the rain occurrence process. Ongoing work focuses on integrating rain intensity into the framework. Looking ahead, we plan to extend the model by incorporating additional meteorological variables, such as temperature and wind speed.

References

P. Ailliot, C. Thompson, and P. Thomson. Space–time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(3): 405–426, 2009. ISSN 1467-9876. doi: 10.1111/j.1467-9876.2008.00654.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2008.00654.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9876.2008.00654.x.

M. Bourotte. *Générateur stochastique de temps multisite basé sur un champ gaussien multivarié*. phdthesis, Université d'Avignon, June 2016. URL https://theses.hal.science/tel-01393617.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the *EM* Algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 39(1):1–22, Sept. 1977. ISSN 1369-7412, 1467-9868. doi: 10.1111/j.2517-6161.1977.tb01600.x. URL https://academic.oup.com/jrsssb/article/39/1/1/7027539.

G. Evin, A.-C. Favre, and B. Hingray. Stochastic generators of multi-site daily temperature: comparison of performances in various applications. *Theoretical and Applied Climatology*, 135(3):811–824, Feb. 2019. ISSN 1434-4483. doi: 10.1007/s00704-018-2404-x. URL https://doi.org/10.1007/s00704-018-2404-x.

E. Gobet, D. Métivier, and S. Parey. Interpretable Seasonal Hidden Markov Model for spatio-temporal stochastic rain generation in France. *Advances in Statistical Climatology, Meteorology and Oceanography (ASCMO), submitted*, 2024.

A. a. C. Klein Tank. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *Int. J. of Climatol., 22, 1441-1453*, 2002.

C. W. Richardson. Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1):182–190, 1981. ISSN 1944-7973. doi: 10.1029/WR017i001p00182. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/WR017i001p00182. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/WR017i001p00182.

A. Touron. *Modélisation multivariée de variables météorologiques*. phdthesis, Université Paris-Saclay, Sept. 2019. URL https://theses.hal.science/tel-02319170.

C. Varin, G. Høst, and Skare. Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics & Data Analysis*, 49(4): 1173–1191, June 2005. ISSN 0167-9473. doi: 10.1016/j.csda.2004.07.021. URL https://www.sciencedirect.com/science/article/pii/S0167947304002403.

Stochastic Rainfall Generator with heavy rainfalls and sustained dry spells

Denis Allard [1] & Antoine Doizé [2]& Philippe Naveau [3] & Olivier Wintenberger [4]

[1] *BioSP, France, denis.allard@inrae.fr*
[2] *LPSM, France, antoine.doize@sorbonne-universite.fr*
[3] *LSCE, France, philippe.naveau@lsce.ipsl.fr*
[4] *LPSM, France, olivier.wintenberger@sorbonne-universite.fr*

Résumé. Dans un contexte de variabilité climatique croissante, l'évaluation précise des phénomènes extrêmes tels que les fortes précipitations et les périodes prolongées de sécheresse ou de pluie est essentielle. En effet ces événements ont des conséquences critiques, tels que des dommages aux infrastructures, les catastrophes agricoles ou même la perturbations de systèmes essentiels tels que le refroidissement des centrales de production électrique. Les modèles climatiques sont devenus des outils incontournables pour étudier l'évolution de ces phénomènes climatiques. Cependant, leur coût computationnel élevé et leur résolution limitée posent des défis importants. Surmonter ces obstacles permettrait de produire des séries de scénarios à des échelles locales (sur une ville) sur de longues périodes (décennies ou siècles), ce qui serait utile pour des études d'impact pour l'aménagement du territoire, l'agriculture ou encore les assurances.

Les générateurs stochastiques météorologiques (Stochastic Weather Generators) offrent une alternative complémentaires qui peuvent aider à dépasser cette barrière computationnelle. Ces outils statistiques exploitent les données météorologiques historiques pour y trouver les tendances statistiques et leur variabilité. Cela permet de simuler des scénarios climatiques sur de longues échelles de temps à un coup computationnel moindre. Cependant, la conception de ces générateurs doit être réalisée avec soin pour produire à la fois des données météorologiques réalistes, mais aussi pour être capable d'extrapoler et de générer des intensités de précipitations extrêmes appropriées ainsi que des périodes prolongées de sécheresse ou de pluie.

Cette présentation décrira un générateur de précipitations journalières capable de produire des scénarios réalistes de précipitations et de périodes sèches, à la fois pour le régime habituel mais aussi dans les extrêmes. Pour ce faire, nous utiliserons une variation des modèles de Markov habituels, en y intégrant des outils de représentation des temps d'attente pour les distributions discrètes (waiting time representation), ainsi que des outils de la théorie des valeurs extrêmes. Cela nous permettra de concevoir un modèle général capable de s'adapter à différents régimes de pluie et permettant une simulation rapide.

Mots-clés. Environnement et statistique, Extrêmes et risques, Statistique appliquée

Abstract. In light of increasing climate variability and its profound impacts, it is critical to accurately assess phenomena such as heavy rainfall and sustained dry spells or rainfalls. These events have severe consequences, including infrastructure damage, agricultural failures, and disruptions of essential systems such as nuclear cooling systems. Climate models have

become an essential tool for projecting these variables in future scenarios. However, their high computational cost and resolution limitations raise significant challenges. Resolving those issues could let us produce numerous local (at a city level) scenarios on long periods (decades or century), which would be helpful for impact studies for land use planning, agriculture or also insurance.

Stochastic weather generators are a complementary tool which can help address this computational barrier. Statistical tools leverage historical data to reproduce climate patterns and their variability, offering a cost-effective means of simulating climate scenarios over long time scales. However, these generators must be designed carefully to produce realistic weather data, and, at the same time must also be able to extrapolate and produce appropriate extreme rainfall intensity, sustained dry spells and rainfalls.

This work presents a daily rainfall generator producing realistic scenarios of rainfall and dry spells. For this we will use a variation of the usual Markov models, by integrating tools of waiting time representation of discrete distributions, and also extreme value analysis. This will let use design a model with tractable inference and quick simulation.

Keywords. Environment and statistics, Extremes and risk, Applied statistics

1 Introduction

In this study, we analyze precipitation data from the Europe Climate Assessment and Dataset (ECAD), focusing on various locations across France. The dataset comprises daily aggregated precipitation data spanning between 50 to 80 years, collected from 50 about different meteorological stations distributed across the country. Our approach involves modeling the precipitation data on a site-by-site basis to capture the unique climatic characteristics of each location.

To model the rainfall data, it is usual to use a Markov chain, modeling rainfall occurrences, and then conditionally modeling rainfall intensities. Such an approach has been introduced by Richardson (1981).

Definition 1 (Markov Model for Rainfall Modeling)**.** *Let $X_0 := x_0$, where $x_0 \in \mathbb{R}_*^+$, and define for any $n \in \mathbb{N}$, $R_n = \mathbb{1}_{X_n>0}$. The model is characterized by the following properties:*

R_n is a two-state homogeneous Markov chain, taking values in $\{0, 1\}$, and thus parametrized by p^0 and p^1 which are respectively the probabilities of leaving state the state 0 and the state 1.

When the current state is dry, i.e. $R_n = 0$, then $X_n = 0$ almost surely.

When the current state is wet, i.e. $R_n = 1$, intensity follows a given distribution

$$(X_n \mid R_n = 1) \sim I_n,$$

where $(I_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. strictly positive random variables, which can, for instance, follow an exponential distribution.

This kind of model has been significantly used and enhanced, for instance, by jointly simulating other variables as in Parlange (2000), or also by extending the model to a multi-site stochastic generator as in Wilks (1998).
However, such a model suffers an issue: sojourn time in a dry spell or in a rainfall are random variable following a geometric distribution. Yet, when looking at recorded duration of dry spells, the bulk of the distribution looks geometric but the tail of the distribution is often heavier: see Figure 1 for an example on 5 specific stations. A direct consequence is that the Markov model suggested in Definition 1 underestimates extreme values of dry spell durations compared to real data (bottom right corner of the figures).This can be harmful for correct risk assessment.

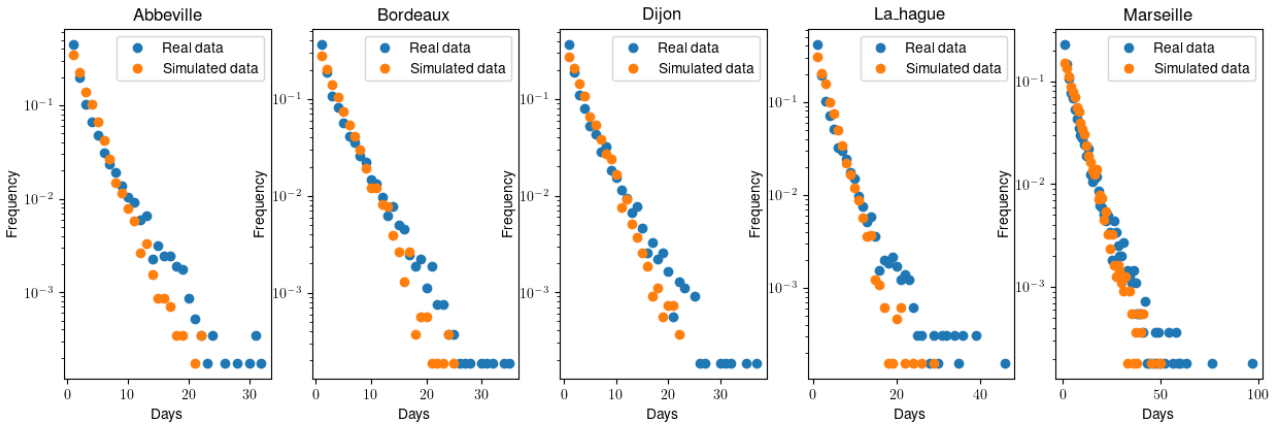


Figure 1: Dry spell distributions, between 1945 and 2024. Blue: recorded values; orange: simulated values according to the Richardson model. Frequencies are in log scale, so that a straight line depicts an exponential distribution.

Intuitively, those geometric sojourn times are directly linked to the fact that the rainfall occurrences $(R_n)_{n\in\mathbb{N}}$ are "only" parametrized by the 2 free parameters of the state transition matrix of a 2-states Markov chain. An idea suggested in Stoner (2022) was to "add" several dry states, which intuitively could stand for "regular dry day state", "severe dry day state". It let them get a more flexible class of distribution for sojourn time and better approximation of the tail of the distribution.

However, even with this approach, the class of sojourn times than can be simulated will still have an exponentially-tailed distributions, which might be too restrictive to describe some stations.

2 Waiting time representation for rainfall occurrence modeling

The idea of this work is to use the waiting time representation of discrete distribution Kozubowski (2025) to be able to reach any discrete distribution for the sojourn time.

Proposition 1 (Kozubowski (2025))**.** *Let T be a discrete random variable on $\mathbb{N}^*$. Let $\{R_n, n \in \mathbb{N}^*\}$ be a sequence of independent Bernoulli trials with success probabilities:*

$$p_n = \mathbb{P}(R_n = 1) = \frac{\mathbb{P}(T = n)}{\mathbb{P}(T \geq n)} = \mathbb{P}(T = n \mid T \geq n),$$

whenever $\mathbb{P}(T \geq n) > 0$, and $p_n = 1$ when $\mathbb{P}(T \geq n) = 0$. Then:

$$T \overset{d}{=} \inf\{n \in \mathbb{N}^* : R_n = 1\}.$$

In this proposition, we can see that T, (which can be seen as the sojourn time in a given state) can be seen as the first success of a sequence of Bernoulli random variables $(R_n)_{n\in\mathbb{N}}$ (which can be seen as the rain occurrence random variables).
Thus, we can design the sequence of success probability of $(R_n)_{n\in\mathbb{N}}$ to match the distribution of any sojourn time T. Here is a scheme for the simulation of such a model:

Definition 2 (Markov Model for rainfall occurrence modeling with waiting time representation)**.** *Let $(p_n^0)_{n\in\mathbb{N}^*}, (p_n^1)_{n\in\mathbb{N}^*}$ be two sequences taking values in $]0,1[$, representing the probability of leaving a given state (wet or dry) after n consecutive days in that state.*

Let $(U_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. standard uniform random variables controlling rain occurrence.

Let us define the Markov chain $(R_n, T_n)_{n\in\mathbb{N}}$, such that R_n is an indicator of rain occurrence on day n, and T_n counts the number of consecutive days that has been spent in the state R_n at day n.

For given initial values $r_0 \in \{0,1\}$ and $t_0 \in \mathbb{N}$, set

$$(R_0, T_0) = (r_0, t_0),$$

and define recursively, for all $n \in \mathbb{N}$:

$$(R_{n+1}, T_{n+1}) =$$

$$\begin{cases} \begin{cases} (1, 1+T_n), & \text{if } U_{n+1} > p^1_{1+T_n}, \\ (0,0), & \text{if } U_{n+1} \leq p^1_{1+T_n}, \end{cases} & \text{if } R_n = 1, \\ \begin{cases} (0, 1+T_n), & \text{if } U_{n+1} > p^0_{1+T_n}, \\ (1,0), & \text{if } U_{n+1} \leq p^0_{1+T_n}, \end{cases} & \text{if } R_n = 0. \end{cases}$$

In our work we are going to focus on discrete extended generalized Pareto distributions as introduced in Naveau (2016). Such distribution let us have control on the behaviour of extreme values, and thus let us simulate extreme values matching real data dynamics.

3 Results

We show in Figure 2 comparison of the duration of dry spell generated with the regular Markov chain described in Definition 1 and with the waiting time representation scheme

describes in Definition 2 after fitting the parameters of an extended-GPD. We can see that for stations such as Marseille with heavy tailed dry spell distributions, simulations with our waiting-time scheme is more effective to reproduce long dry spells.

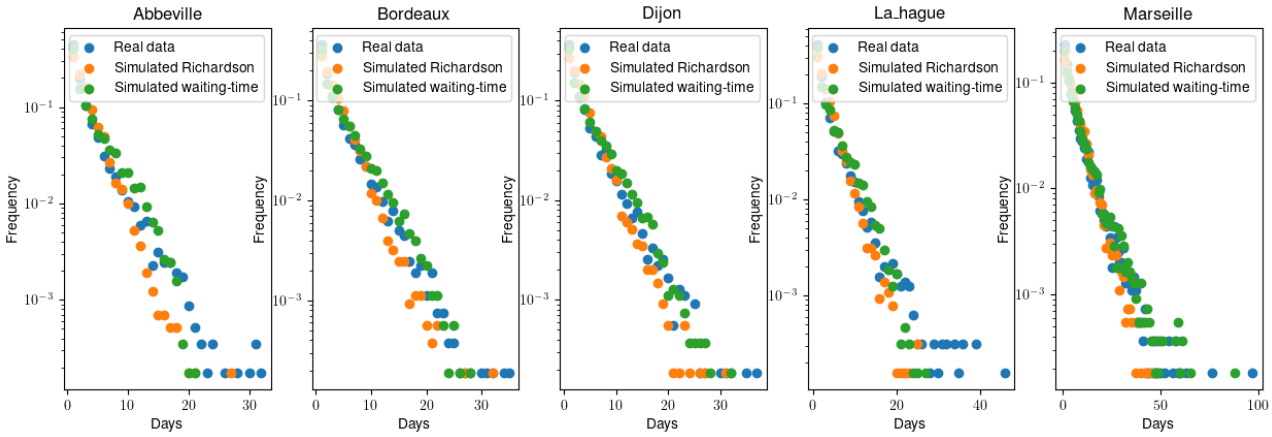


Figure 2: Dry spell distributions, between 1945 and 2024. Blue: recorded values; orange: simulated values according to the Richardson model; green: simulated values according to a waiting time representation model. Frequencies are in log scale, so that a straight line depicts an exponential distribution.

In the presentation we will also show how to model the rain intensity to reproduce both the bulk and the tail of its distribution.

Bibliographie

Kozubowski T. J., Mlynarczyk D., Panorska A. K. (2025), Waiting time representation of discrete distributions, *Statistics and Probability Letters*,

Naveau P., Huser R., Ribereau P., Hannart, A. (2016), Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection, *Water Resources Research*, 52

Parlange, M. B. (2000), An Extended Version of the Richardson Model for Simulating Daily Weather Variables, *Journal of Applied Meteorology and Climatology*, 39, pp. 610-622.

Richardson, C. W. (1981),Stochastic simulation of daily precipitation, temperature, and solar radiation, *Water Resources Research*, 17, pp. 182-190.

Stoner, R. O., Ecconomou, T. (2020), An advanced hidden Markov model for hourly rainfall time series, *Computational Statistics and Data Analysis*, 152.

Wilks, D. S. (1998),Multisite generalization of a daily stochastic precipitation generation model, *Journal of Hydrology*, 210, pp. 178-191.

Distribution composite basée sur le test de Berk-Jones pour la correction de biais saisonnière de précipitations journalières

Philippe EAR[1], Elena DI BERNARDINO[2], Thomas LALOË[3]

[1] *Hydroclimat, LJAD, Université Côte d'Azur, France, philippe.ear@hydroclimat.com*
[2] *LJAD, Université Côte d'Azur, France, elena.di_bernardino@unice.fr*
[3] *LJAD, Université Côte d'Azur, France, thomas.laloe@unice.fr*

Résumé. Pour que les modèles d'impact estiment correctement les evènements extrêmes, il est nécessaire de disposer de données journalières de précipitations qui soient hautement résolues et précises. Le test de goodness-of-fit de Berk-Jones diffère des autres tests classiques en prenant en compte la significativité statistique des écarts entre les quantiles empiriques et théoriques. Il permet une meilleure détection des écarts, en particulier pour les données à queue lourde. Dans une étude précédente, le Stitch-BJ a été présenté comme une distribution semi-paramétrique basée sur le test de Berk-Jones pénalisé qui permet une jonction automatique et personnalisée de distributions paramétriques et empirique. Ici, nous validons le modèle dans un contexte d'une correction de biais avec une séparation saisonnière, d'une correction de la probabilité de jours secs et d'une validation out-of-sample, en se comparant à d'autres modèles tels que les distributions Gamma, Exponentiated Weibull, Extended Generalized Pareto ou empirique. En outre, une étude simulée sur la version pénalisée du test de Berk-Jones a été réalisée afin de renforcer la confiance dans son applicabilité. Le Stitch-BJ a été en mesure d'obtenir des résultats aussi bons, voire meilleurs, que tous les autres modèles sur l'ensemble de validation, y compris la distribution empirique, qui est souvent utilisée en raison de sa robustesse.

Mots-clés. Correction de biais, Distribution paramétrique, Goodness-of-fit, Modélisation de précipitation, Théorie des valeurs extrêmes

Abstract. Highly resoluted and accurate daily precipitation data are required for impact models to perform adequately and to correctly measure high-risk events' impact. The Berk-Jones (BJ) goodness-of-fit statistical test differs from other classical goodness-of-fit test by considering the statistically significance of deviations between empirical and theoretical quantiles. It allows us for a better detection of deviations, especially in heavy-tailed data. In a previous study, the co-called *Stitch-BJ method* was introduced as a semiparametric distribution based on a penalized BJ statistical test, which allows for an automatic and personalized splicing of parametric and nonparametric distributions. Here, we further validate the model in a bias-correction context with seasonal separation, correction of dry days probability and out-of-sample validation, against other models such as the Gamma, Exponentiated Weibull, Extended Generalized Pareto or empirical distribution. Moreover, simulation studies on the penalized versions of the Berk-Jones test were provided. The Stitch-BJ was able to consistently perform as good, even better, than all the other models over the validation set, including the empirical distribution, which is often used due to its robustness.

Keywords. Bias correction, Extreme value theory, Goodness-of-Fit, Parametric distribution, Precipitation modelling

1 Introduction

Highly resoluted and unbiased daily precipitation data are often needed for hydrological model and impact models (see, e.g., (Casanueva et al., 2016)). However, most readily available data are either too coarse, biased or have limited availability in time. Increasing resolution and removing bias can then be done by performing bias correction (Cannon, Sobie, and Murdock, 2015), the most common technique being the empirical quantile mapping. However, using an empirical distribution prevents from extrapolating to new extremes, which in result may underestimate the frequency or intensity of extreme events, especially in a climate change context. The use of parametric distributions have been explored in the past (Naveau et al., 2016; Mamalakis et al., 2017) with promising results, however classical parametric distributions often fail to correctly model daily precipitation over large and highly resoluted study areas. In the past few years, new parametric and semi-parametric models have been studied to bridge the gap between classical parametric distributions with their abilities to extrapolate extreme values, and the robustness and adaptability of empirical distributions. Such models may be fully parametric, with the presence of a stitch or not (see, e.g., Naveau et al., 2016; Mamalakis et al., 2017), or semi-parametric (see, e.g., Holthuijzen et al., 2022; Ear et al., 2024).

Here, we focus on the last semi-parametric model recently introduced in Ear et al. (2024) and referred as the *Stitch-BJ model*. A seasonal bias correction using quantile mapping, including dry days probabilities correction using the Singularity Stochastic Removal method from Vrac, Noël, and Vautard (2016) is performed, with comparisons with competitor distributions. Moreover, a simulation study is conducted on the proposed penalized Berk-Jones test to validate the gain of power without compromising the statistical significance of the result.

2 Statistical methodology

2.1 Quantile mapping for the considered data

For this study, two European reanalysis data sets produced by the European Center for Medium-Range Weather Forecast (ECMWF) will be used: the ERA5-L reanalysis (resolution of $0.1° \times 0.1°$) as model data to-be-corrected, and the CERRA-L reanalysis (resolution of $5.5 \times 5.5km$) as reference data. We selected a common period for both reanalyses, starting from 01/01/1985 to 31/12/2020, with a separation into training and validation period : respectively 01/01/1985 - 31/12/2009 and 01/01/2010 - 31/12/2020. Both reanalyses were separated using the meteorological season (DJF, MAM, JJA, SON) to take into account intra annual non-stationarity property of the considered time-series.

The quantile mapping is a cumulative distribution function (CDF) based transformation. Let $F_{ref,mod}$ and $F_{ref,obs}$ respectively be the CDF of a given location for the ERA5-L and CERRA-L data for the training period, and $x_{fut,mod}$ the ERA5-L data for the validation period. Then the corrected model data is :

$$\hat{x}^{corr}_{fut,mod} = F^{-1}_{ref,obs}(F_{ref,mod}(x_{fut,mod}))$$

The quantile mapping requires to set distributional models for F. In this study, in order to assess the bias-correction performance, we compare our *Stitch-BJ model* previously

introduced in Ear et al., 2024 to the following models : Empirical distribution, Extended Generalized Pareto (EGP) distribution (see Naveau et al., 2016), Exponentiated Weibull (ExpW) distribution and Gamma distribution.

2.2 A penalized Berk-Jones goodness-of-fit test

The Berk-Jones test is based on the Berk-Jones statistic which measures the statistical significance of the CDFs deviations instead of their absolute values. It has been claimed to dominate more classical statistics, especially when it comes to heavy-tailed distributions. We refer to Moscovich, Nadler, and Spiegelman, 2016 for most of the results and discussion about the Berk Jones test and statistic.

• Let $\tilde{x} := \{\tilde{x}_1, ..., \tilde{x}_n\}$ be an $i.i.d$ sample from a random variable X. Let X follow a continuous cumulative distribution function F. Then, under the null hypothesis $\mathcal{H}_0 : X \sim F$. The alternative hypothesis will be $\mathcal{H}_1 : X \not\sim F$.
• Let $x := \{x_{(1)}, ..., x_{(n)}\}$ be empirical quantiles of the sample $\tilde{x}$, with $x_{(1)} < x_{(2)} < ... < x_{(n)}$ which are the order statistics of x (we can safely assume that none of the $i.i.d$ samples are equal).
• We also denote the quantile of a distribution F corresponding to the i-th ordered quantile as $F^{-1}(\frac{i-1}{n}) = q_{(i),F}$, with F the candidate distribution fitted on x. Let $q := \{q_{(1),F}, ..., q_{(n),F}\}$.

Let consider $u_{i,F} = F(x_{(i)})$ for $i = 1, \dots, n$ an $i.i.d$ sample from a $U[0, 1]$ distribution and the corresponding order statistics $u_{i,F}$. We know that under $\mathcal{H}_0, u_{i,F}$ is a random sample from u_i which has a known distribution: $u_i \sim \text{Beta}(i, n - i + 1)$, $for\, i = 1, \dots, n$.

The BJ test then computes the following p-values: $p^*_{i,F} := \mathbb{P}(u_i < u_{i,F})$. For each i, we can then define:

$$p_{i,F} = \min(p^*_{i,F}, 1 - p^*_{i,F}). \tag{1}$$

The M_n statistic is defined by:

$$M_n := \min_{1 \le i \le n} p_{i,F} \in [0, 1], \tag{2}$$

where small values of M_n describe a bad fit of the candidate distribution F. The p-value associated to α, denoted $p_{\alpha,n}$, is given by $\mathbb{P}(M_n \le p_{\alpha,n}) = \alpha$. In the modified Berk-Jones test introduced in Ear et al. (2024), the idea is to penalized the $p_{(i)}$ from Equation (1) and thus the M_n from Equation (2) by the absolute error between the theoretical and the empirical quantile for the given rank i. We introduce the levels k_i:

$$k_i = \frac{p_{(i)}}{\max(1, |F^{-1}(\frac{i-1}{n}) - x_{(i)}|)} \tag{3}$$

The rest of the test is done similarly as the BJ test. Since $k_i \le p_{(i)}$, the increase in power is clear if we use the same rejection threshold. However, this could impact the significance of the test and reduce its usability.

Since the new statistic is dependent on both the distribution F and the sample x, obtaining theoretical guarantees on the power, significance and consistency of the test still requires further work. However, a numerical study can be conducted to increase our confidence in its practical usability (see Section 3.1 below).

3 Results

3.1 Numerical study on penalized Berk-Jones test

For all results of the numerical study, distributions were fitted using the maximum likelihood estimator and all experiments were repeated 1000 times, for each ($\mathcal{H}_0$/simulated distribution) couple, and for each sample size with significance level $\alpha = 5\%$.

Table 1: Table of scenarios with the considered distributions for the numerical study

Scenario	**1**	**2**	**3**	**4**	**5**
$\mathcal{H}_0$	EGP	ExpW	EGP	ExpW	EGP
Simulated	EGP	ExpW	ExpW	EGP	Stitch EGP-ExpW

For each experiment, we simulated samples from a given distribution that was either the same distribution as $\mathcal{H}_0$, or a different one among the previously listed models (see Table 1 for a description of the different scenarios used). Then we fitted the $\mathcal{H}_0$ model to the simulated data using the maximum likelihood estimator before applying goodness-of-fit tests.

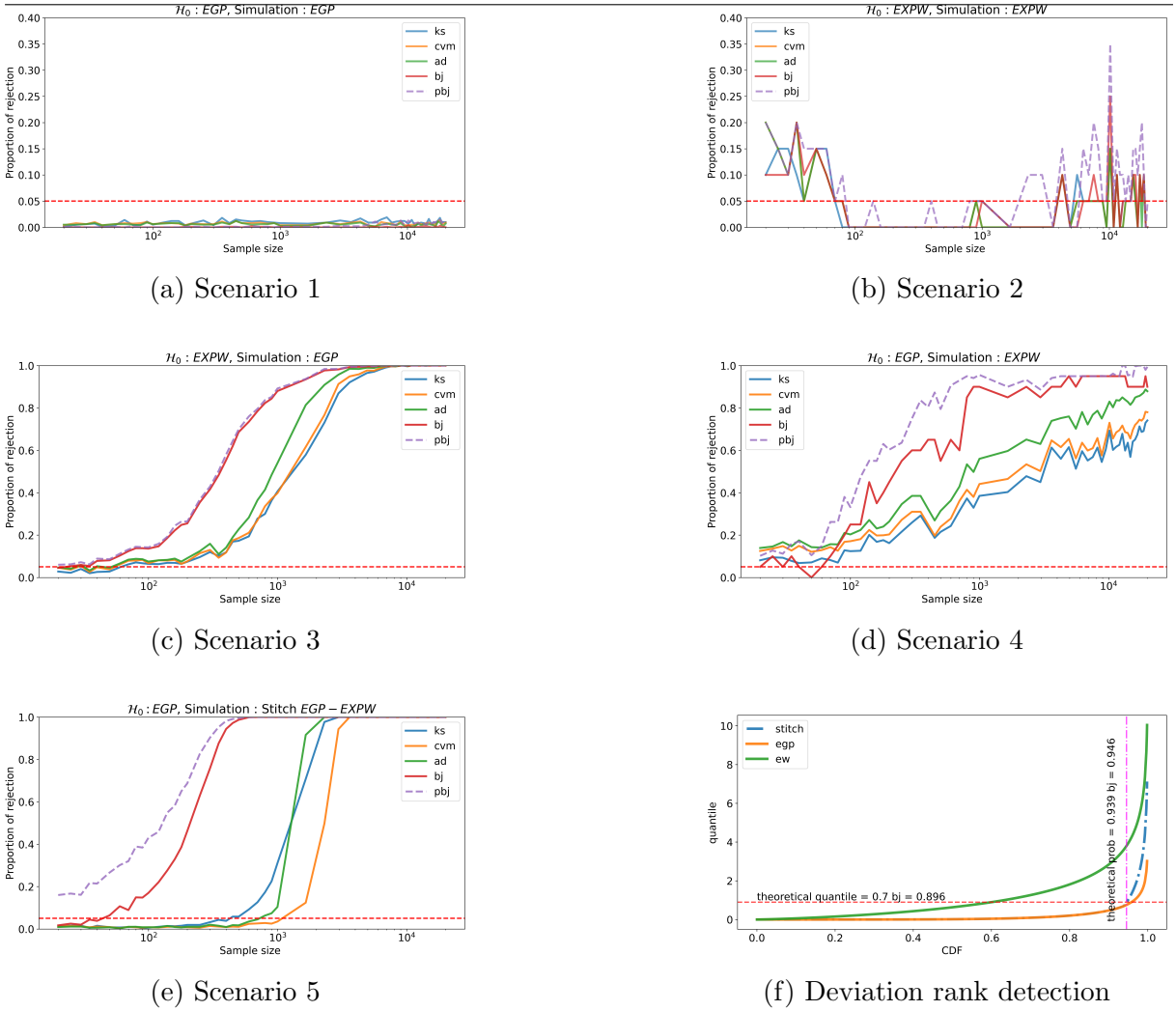
In Figure 1, we compare the performance of the (non-penalized) Berk-Jones test (BJ), against the Penalized Berk-Jones test (PBJ), the Kolmogorov-Smirnov test (KS), the Cramér-von Mises test (CVM) and the Anderson-Darling (AD) test.

For Scenario 1, all goodness-of-fit tests performed equally well, with less than 5% of false rejection for all sample sizes. Under Scenario 2, for small sample size (less than 50), all tests showed a higher rejection rate than the $\alpha = 5\%$ significance level chosen. For a large sample size (greater than 5000), the BJ and PBJ tests and occasionally the others, showed higher significance levels than expected (see Figure 1 panels a and b). However, when the simulated distribution is different from $\mathcal{H}_0$ (Scenario 3, 4 and 5), the BJ and PBJ tests became equally or more powerful than all other alternatives for the vast majority of sample size, with also a much quicker convergence with increasing sample size (see panels c, d, e in Figure 1). The PBJ test is, by definition, always more powerful than the standard BJ test. Panel (f) shows how the PBJ test can be used to detect the deviation rank between two distributions, using $p_{(i)}$ and k_i defined in Equations (1) and (3).

3.2 Seasonal bias-correction results

For precipitation modeling, a single parametric distribution may not always be able to model a large study area due to the spatial heterogeneity and the intra-annual non-stationarity. Here, we used the PBJ tests (see Ear et al. (2024) for details about the specific algorithm) to detect misfitted locations and construct a stitched distribution to better match the location's data, using a possible mixture of the candidate distributions introduced in Section 2.1.

In Figure 2, we have for each location the description of the distribution used when fitted on CERRA-L data over the DJF (panel a) and MAM (panel b) seasons. If only one distribution is given, then it is used fully to model the data. With two distributions, the second one is used to model the upper tail. With three distributions, the first one models the lower tail, while the third one models the upper one.

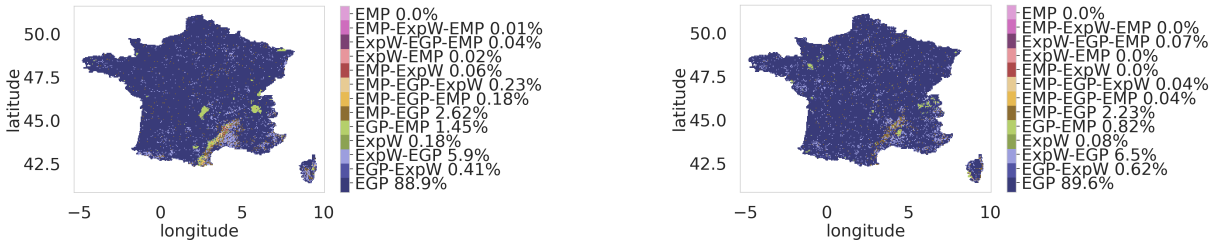


(a) Scenario 1 (b) Scenario 2

(c) Scenario 3 (d) Scenario 4

(e) Scenario 5 (f) Deviation rank detection

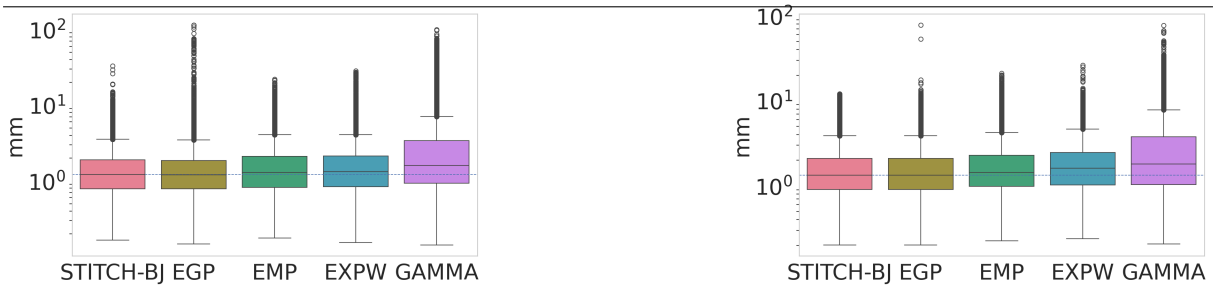
Figure 1: Panels a-e: Comparison of the power between KS, CVM, AD, BJ and PBJ test for $\alpha = 5\%$ for Scenarios 1-4. Panel f: An illustration on how an algorithm based on the PBJ test can be used to detect the deviation's starting from a certain rank i between two distributions.



(a) December-January-February (DJF) (b) March-April-May (MAM)

Figure 2: Stitch classes and associated proportions fitted on CERRA-L 01/01/1985-31/12/2009.

Finally, in Figure 3, we can see that the Stitch-BJ model is able to remove extreme outliers with respect to the EGP distribution, while keeping the median RMSE error similar to the empirical model one, for both displayed seasons. This shows that we effectively targeted locations where the EGP produced high errors and we replaced it



(a) RMSE boxplot for DJF

(b) RMSE boxplot for MAM

Figure 3: Boxplots of RMSE for the bias corrected ERA5-L using several models. Note that the y-axis is in log scale.

with either the empirical or ExpW distributions. This allows us to have an automatic procedure that produces mostly parametric distributions over a large study area with similar performance to the empirical model, but allowing for possible extrapolations of extremes.

References

Cannon, A. J., S. R. Sobie, and T. Q. Murdock (2015). "Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes?" *Journal of Climate* 28.17, pp. 6938–6959.

Casanueva, A. et al. (2016). "Daily precipitation statistics in a EURO-CORDEX RCM ensemble: added value of raw and bias-corrected high-resolution simulations". *Climate Dynamics* 47.3, pp. 719–737.

Ear, P. et al. (2024). *A semi-parametric distribution stitch based on the Berk-Jones test for French daily precipitation bias correction*. Under review.

Holthuijzen, M. et al. (2022). "Robust bias-correction of precipitation extremes using a novel hybrid empirical quantile-mapping method: Advantages of a linear correction for extremes". *Theoretical and Applied Climatology* 149.1-2, pp. 863–882.

Mamalakis, A. et al. (2017). "A parametric approach for simultaneous bias correction and high-resolution downscaling of climate model rainfall". *Water Resources Research* 53.3, pp. 2149–2170.

Moscovich, A., B. Nadler, and C. Spiegelman (2016). "On the exact Berk-Jones statistics and their p-value calculation". *Electronic Journal of Statistics* 10.2.

Naveau, P. et al. (2016). "Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection". *Water Resources Research* 52.4, pp. 2753–2769.

Vrac, M., T. Noël, and R. Vautard (2016). "Bias correction of precipitation through Singularity Stochastic Removal: Because occurrences matter". *Journal of Geophysical Research: Atmospheres* 121.10, pp. 5237–5258.

Extrêmes et changement climatique

Dynamique des événements extrêmes et processus ponctuels

Antoine Heranval[1] & Denis Allard[2] & Édith Gabriel[3] & Thomas Opitz[4]

[1] *INRAE, BioSP, France – antoine.heranval@inrae.fr*
[2] *INRAE, BioSP, France – denis.allard@inrae.fr*
[3] *INRAE, BioSP, France – edith.gabriel@inrae.fr*
[4] *INRAE, BioSP, France – thomas.opitz@inrae.fr*

Résumé. Dans un contexte de changement climatique, il est devenu essentiel de comprendre et de modéliser les événements météorologiques extrêmes. Cette étude propose une approche innovante reposant sur les processus ponctuels marqués pour analyser la dynamique spatio-temporelle des événements extrêmes. En utilisant les données ERA5 du Copernicus Data Store, nous examinons les précipitations extrêmes et les anomalies de température à l'échelle des régions NUTS européennes. Notre méthodologie associe l'analyse des processus ponctuels à la théorie des valeurs extrêmes, permettant ainsi une caractérisation fine des dépendances spatiales et temporelles. Les événements sont définis par des dépassements de seuils et sont étudiés à travers les fonctions de corrélation croisée K. Les premières analyses révèlent une forte corrélation spatiale des précipitations extrêmes entre régions adjacentes, tandis que les régions éloignées montrent une indépendance marquée. Les anomalies de température, qu'elles soient chaudes ou froides, présentent une corrélation négative, suggérant un effet répulsif. Cette approche ouvre des perspectives prometteuses pour l'analyse des risques climatiques complexes et l'évaluation des impacts du changement climatique.

Mots-clés. Processus ponctuels marqués, Événements extrêmes, Théorie des valeurs extrêmes, Risques climatiques, Analyse spatio-temporelle

Abstract. In the context of climate change, understanding and modeling extreme weather events has become crucial. This study proposes an innovative approach based on marked point processes to analyze the spatio-temporal dynamics of extreme events. Using ERA5 data from the Copernicus Data Store, we examine extreme precipitation and temperature anomalies at the NUTS regional level in Europe. Our methodology combines point process analysis with extreme value theory, allowing for a detailed characterization of spatial and temporal dependencies. Events are defined by threshold exceedances and analyzed through cross-correlation functions K. Preliminary analyses reveal a strong spatial correlation of extreme precipitation between neighboring regions, while distant regions show marked independence. Both hot and cold temperature anomalies exhibit a negative correlation, suggesting a repulsive effect. This approach opens promising perspectives for analyzing compound climate risks and assessing the impacts of climate change.

Keywords. Marked point processes, Extreme events, Extreme value theory, Climate risks, Spatio-temporal analysis

1 Introduction and Motivation

Climate change has led to an increased focus on understanding and modeling extreme weather events. The presentation, addresses this critical challenge by applying marked point processes to analyze extreme events. The primary objective is to transform climate model "big data" into actionable knowledge about extreme events and climate change, utilizing methods from point processes, extreme value theory, and spatial graphs.

2 Theoretical Framework: Marked Point Processes

Marked Point Processes provide a mathematical framework for analyzing discretely observed stochastic events (Reinhart, 2018). In this framework, each event is characterized by a point representing coordinates in time, space, or space-time, and a mark consisting of one or more numerical variables such as duration, size, or intensity of the event. This approach can be applied to various contexts, from spatial processes like urban heat islands to temporal processes such as rainfall bursts during storms, and spatio-temporal processes like wildfire occurrences.

In this context, we can consider two key representations for temporal point processes : the first is the counting process, defined as $N(t) = \sum_i \mathbb{I}(t_i \leq t)$, and the second is the cumulative mark process, expressed as $M(t) = \sum_i m_i \mathbb{I}(t_i \leq t)$. These processes are characterized by their intensity function λ, which represents the mean number of events in a time interval (stationary in this case):

$$\lambda = \frac{1}{T}\mathbb{E}\left[\#(t_i \in [0, T])\right]$$

The correlation structure is captured by the K-function:

$$K(r) = \frac{1}{\lambda}\mathbb{E}\left[\#\{t_i \in [t_0 - r, t_0 + r]\} \mid \text{event at } t_0\right]$$

For analyzing multiple point processes, Cebrián et al. (2020) introduced a cross-K function:

$$\hat{K}_{xy}(r) = \frac{1}{T}\sum_{t_i \in N_x}\sum_{s_j \in N_y}\frac{I(|t_i - s_j| \leq r)}{\lambda_x \lambda_y}$$

3 Data Analysis and Implementation

The study utilizes ERA5 hourly data from the Copernicus Data Store, (Hersbach et al., 2023). The analysis incorporates multiple variables including wind components, 2m temperature, total precipitation, sea level pressure, soil moisture, and cloud cover. The quality and

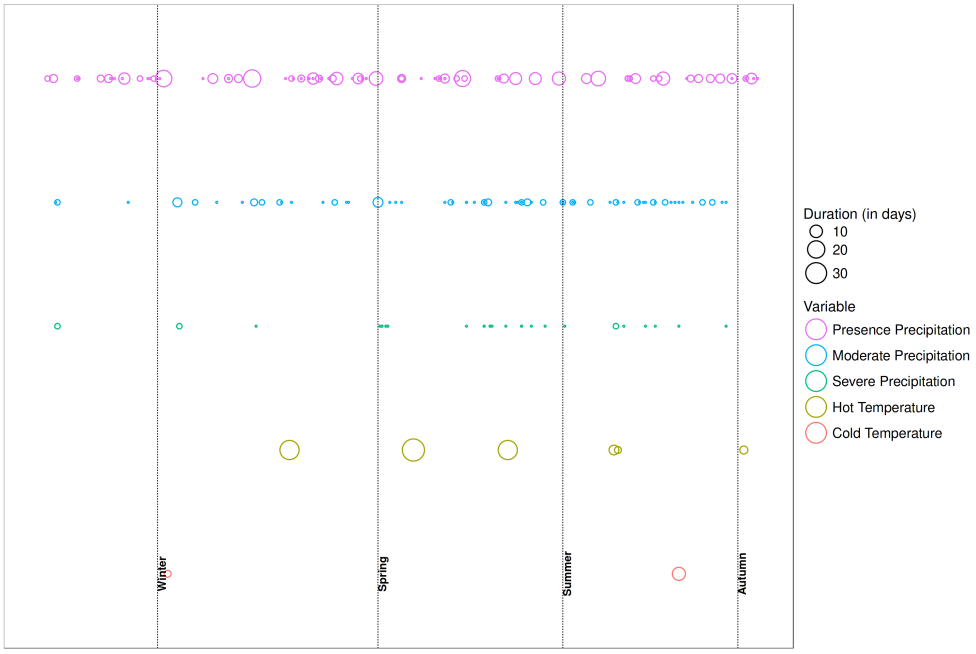


Figure 1: Temporal Marked Point Processes for Languedoc-Roussillon in 2022, showing the distribution of extreme events and their temporal clustering patterns.

reliability of ERA5 precipitation data for climate monitoring has been validated by (Lavers et al., 2022).

The definition of extreme events follows a systematic approach. For precipitation, events are classified based on daily accumulation, with extreme events defined as those exceeding the 99th percentile. Temperature extremes are identified through anomalies from the seasonal historical mean, with hot events corresponding to anomalies above the 95th percentile and cold events to those below the 5th percentile.

The implementation of the first analysis relies on established statistical software packages. The IndTestPP package (Cebrián and Asín, 2021) is used for testing independence between point processes, while spatial analysis is conducted using the stpp package (Gabriel et al., 2013).

A key aspect of the study involves analyzing spatial dependencies between extreme events in neighboring regions using the cross K-function methodology. This statistical tool, developed by (Cebrián et al., 2020), provides valuable insights into the spatial and temporal relationships between point processes.

The analysis reveals significant spatial dependency patterns between neighboring regions. As shown in Figure 2, the cross K-function for severe precipitation events between the Languedoc-Roussillon and Provence-Alpes-Côte d'Azur regions demonstrates strong positive correlation, particularly at short time scales. The function exceeds the 95% confidence bounds, indicating clustering behavior that cannot be explained by random chance. This finding suggests that extreme precipitation events in these Mediterranean regions tend to

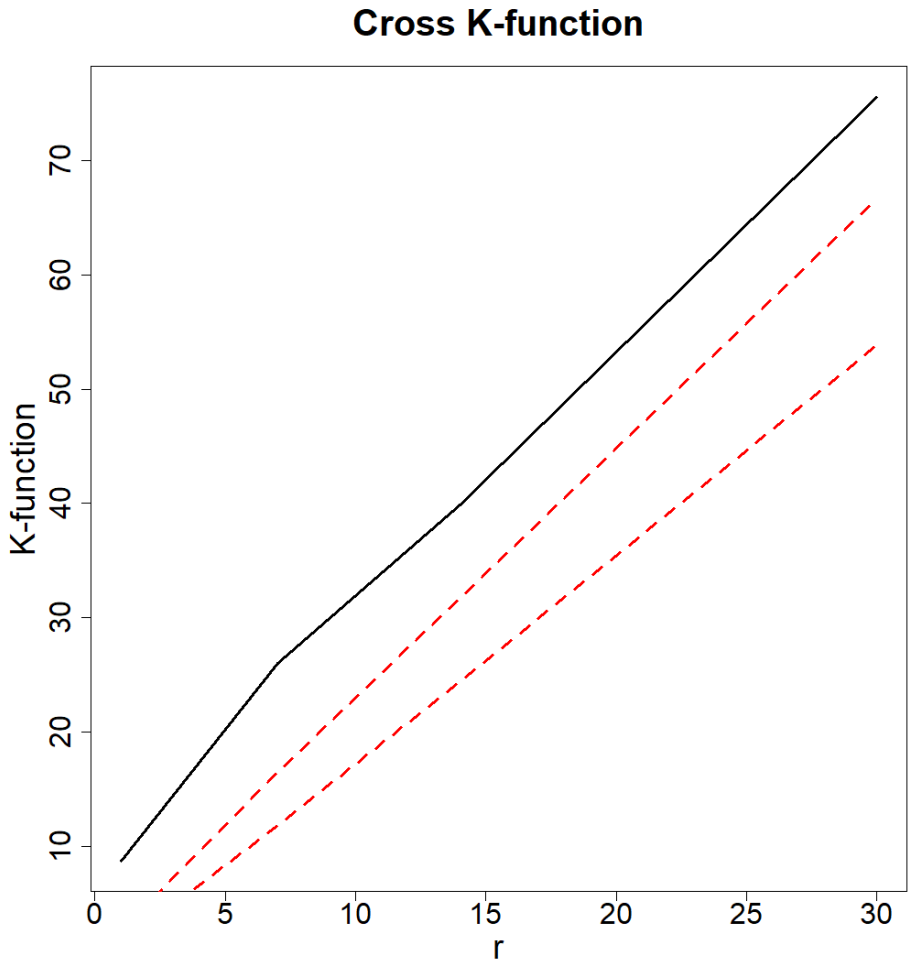


Figure 2: Cross K-function analysis for severe precipitation between Languedoc-Roussillon and Provence-Alpes-Côte d'Azur regions. The dashed lines represent the 95% confidence interval under the independence hypothesis.

occur in temporal proximity, as we can exepect since they share the same meteorological systems affecting both areas simultaneously.

This spatial dependency structure has important implications for risk assessment and management. The clustering of extreme events across neighboring regions suggests that risk mitigation strategies should consider the potential for synchronized impacts across multiple administrative boundaries. Furthermore, the temporal scale of these correlations provides valuable information for early warning systems and emergency response planning.

The strength of the observed spatial correlations decreases with distance between regions, as demonstrated by comparison with more distant areas. This spatial decay in correlation provides insights into the typical spatial extent of extreme weather systems and helps define the relevant scales for risk assessment.

4 Key Findings and Discussion

The analysis reveals several significant patterns in extreme event occurrence. Strong spatial correlation exists between neighboring NUTS regions for precipitation events, while distant regions show minimal correlation. This spatial dependence structure aligns with meteorological understanding of weather systems and their spatial extent. Temperature anomalies (hot and cold events) demonstrate a negative correlation, suggesting a repulsive effect that merits further investigation. Of particular interest is the relationship between cold events and severe precipitation, which shows slight positive correlation at short time scales.

5 Future Directions

The research opens several promising directions for future investigation. These include the extension to other perils, particularly convective events, and the investigation of long-term climate projections and trends. The development of compound extreme events analysis and the application of conditional extremes models represent particularly promising areas for future research. The methodology developed here provides a robust framework for analyzing spatial and temporal patterns of extreme weather events, contributing to climate change adaptation and risk assessment strategies.

References

Cebrián, A. C., Abaurrea, J., and Asín, J. (2020). Testing independence between two non-homogeneous point processes in time. *Journal of Statistical Computation and Simulation*, 90(16):2878–2901.

Cebrián, A. C. and Asín, J. (2021). Analyzing dependence between point processes in time using indtestpp. *The R Journal*, 13:444–460. https://rjournal.github.io/.

Gabriel, E., Rowlingson, B. S., and Diggle, P. J. (2013). stpp: an r package for plotting, simulating and analyzing spatio-temporal point patterns. *Journal of Statistical Software*, 53:1–29.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, , Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N. (2023). Era5 hourly data on single levels from 1940 to present. Accessed on 10-09-2024.

Lavers, D. A., Simmons, A., Vamborg, F., and Rodwell, M. J. (2022). An evaluation of era5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society*, 148(748):3152–3165.

Reinhart, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318.

Exploring Climate Change Through the Lens of Records Theory

Jinane Jouni[1], Anis Hoayek[2], Aline Mefleh[3] & Mireille Batton-Hubert[4]

[1] *Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F - 42023 Saint-Etienne France, jinane.jouni@hotmail.com*
[2] *Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F - 42023 Saint-Etienne France, anis.hoayek@emse.fr*
[3] *Department of Mathematics, Holy Spirit University of Kaslik, Jounieh, Lebanon, aline.mefleh@usek.edu.lb*
[4] *Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F - 42023 Saint-Etienne France, batton@emse.fr*

Résumé. Le phénomène mondial du changement climatique est devenu l'un des défis les plus urgents de notre époque. Dans ce contexte complexe, la communauté scientifique cherche des méthodologies innovantes pour quantifier et comprendre l'occurrence et la gravité du changement climatique. La théorie de records en statistique des extrêmes offre un cadre prometteur pour détecter et vérifier le changement climatique. Notre recherche couvre plusieurs lieux géographiques et utilise une approche de notation univariée pour analyser les tendances de température et de précipitations, fournissant ainsi des informations précieuses sur l'occurrence, la fréquence et l'intensité des événements climatiques extrêmes. Nos résultats révèlent un schéma de réchauffement global, avec des scores de température en baisse au milieu du siècle avant d'augmenter brusquement, dépassant souvent les niveaux précédents. Les scores de précipitations varient selon les régions, certaines connaissant des conditions plus sèches tandis que d'autres font face à une augmentation des précipitations. Ces résultats mettent en évidence des extrêmes climatiques changeants, soulignant l'urgence d'analyser le changement climatique.

Mots-clés. Changement climatique, Théorie de records, Notation statistique, Température, Précipitation

Abstract. The global phenomenon of climate change has become one of the most pressing challenges of our time. Amidst this complex landscape, the scientific community seeks innovative methodologies to quantify and understand the occurrence and severity of climate change. The theory of records in extreme value theory provides a promising framework for detecting and verifying climate change. This research spans multiple geographic locations, employing a univariate scoring approach to analyze temperature and precipitation trends, and providing valuable insights into the occurrence, frequency, and intensity of extreme climate events. Our findings reveal a global warming pattern, with temperature scores declining mid-century before rising sharply, often surpassing previous levels. Precipitation scores vary across regions, with some experiencing drier conditions while others face increased rainfall. These results highlight shifting climate extremes, reinforcing the urgency of climate change analysis.

Keywords. Climate change, Records theory, Scoring system, Temperature, Precipitation

1 Introduction

Climate change is one of the most urgent global challenges, with its environmental and societal impacts becoming increasingly evident. Over the past several decades, scientific consensus has established that human activities are the primary drivers of climate change, with mean global surface temperature increasing and reaching 1.1°C above pre-industrial levels in 2011-2020 (IPCC, 2023). However, variability and extremes are critical to understanding the full impact of climate change beyond the global mean value (Schär et al., 2004), especially that the evolution of extreme temperatures does not always follow that of the mean (Castillo-Mateo et al., 2023). Understanding the manifestations and trends of climate change is crucial for effective mitigation strategies and informed decision-making.

In parallel, record theory provides a powerful framework for analyzing rare extreme events by examining the statistical behavior of record values in time-series data. In a time series $\{X_t\}_{1\leq t\leq T}$, where T is the present time, a record $R_n = X_{L_n}$ occurs at time L_n when X_{L_n} exceeds all previous values, that is, $X_{L_n} > \max\{X_1, ..., X_{L_n-1}\}$.

Four key models have been developed to characterize record-breaking events. The most fundamental independent and identically distributed model (*i.i.d.*) assumes records occur independently, rarely, and symmetrically (the same number of max and min records, forward and backward records) in a stationary climate context (Chandler, 1952; Arnold et al., 1998). The Discrete-Time Random Walk Model (DTRW), discussed by Majumdar and Ziff (2008), accounts for random fluctuations and increments, making it useful for short-term weather variability chaotic conditions. The Linear Drift Model (LDM) introduces a constant linear drift influencing the occurrence of records and representing a steady warming or cooling trend rather than random fluctuations steady. Lastly, the Yang-Nevzorov Model increases the likelihood of records over time as the number of record-candidate observations increases exponentially, representing scenarios with accelerated warming (Nevzorov, 1990). Table A.1 in the appendix summarizes each model.

This study aims to bridge the gap between theoretical record-breaking frameworks and their practical applications in analyzing, evaluating, and comparing climate data across different regions by exploring patterns of extreme events. The findings will contribute to a more detailed understanding of climate change's regional impacts and provide new

statistical method for future studies. This study is organized into four sections: Section 2 reviews previous applications of record theory in climate studies, particularly for temperature and precipitation. Section 3 details the methodology, including feature selection, categorization and scoring system. Section 4 presents the results for six geographical stations, while Section 5 synthesizes the findings, discusses implications, and proposes future research directions.

2 Literature Review

Several studies have applied record theory to analyze temperature and precipitation trends in various regions. For example, Benestad (2003) analyzed the incidence of temperature and precipitation events (moving forward and backward in time) in multiple dependent stations of Northern Europe (1895–1999), and found a significant increase in high-temperature records moving forward compared to what is expected in a stationary *i.i.d.* model. Similarly, Meehl et al. (2009) observed a 2:1 ratio of high-to-low records in U.S. temperature data (1950–2006) since the late 1970s,with projections suggesting ratios as high as 50:1 by the end of the 21st century. In Europe, Elguindi et al. (2013) identified a rising trend in high-temperature records and stable to declining low-temperature records since the 1980s, consistent with regional warming patterns.

Studies have also explored the influence of trends and variability on record-breaking events. Newman et al. (2010) demonstrated that linear trends significantly alter the ratio of high-to-low record-breaking temperatures beyond 1:1, as seen at the Mauna Loa Observatory, where minimum temperature records exhibited a stronger warming trend than maximum records. Using the Linear Drift Model, Redner and Petersen (2006) and Wergen and Krug (2010) fitted a Gaussian distribution for daily temperature data in Europe and the U.S., observing a consistent increase in mean temperatures with stable variance, predicting ongoing rises in high-temperature records alongside declines in low-temperature records.

Overall, most previous studies anchor their comparisons against the *i.i.d.* model without attempting to fit specific record models to the data. Moreover, these studies predominantly focus on one statistic, such as record counts, max-to-min ratios, or linear trends to conclude to global warming. This study seeks to address these limitations by integrating additional variables and applying a comprehensive scoring system to analyze climate trends more holistically.

3 Proposed Methodology

The proposed methodology employs a uni-dimensional framework to analyze trends in climate measures—primarily temperature and precipitation—by constructing a new time series, $\{S_t\}_{1\leq t\leq w}$ (w will designate the number of subset series or so-called moving windows). This series is derived using a weighted scoring system based on a set of predefined features. This section elaborates on the methodology in three steps: feature development, feature selection and weighting, and the scoring framework.

1. Feature Development

   Based on record theory and statistical principles, a comprehensive set of variables was developed aiming to measure or assess different aspects of records. These variables were grouped into three factors—Frequency, Intensity, and Trend—each providing distinct insights into the behavior of records.

   (a) **Frequency-based features** quantify the occurrence and distribution of records within a given time frame. Examples include the record count, i.e the number of record-breaking observations within a time frame. Another variable is the model-based score where higher values are assigned to models with greater probabilities of generating records (e.g., *i.i.d.*: 1, DTRW: 2, LDM/Yang: 3; see Figure 1). Model fitting is validated using statistical tests such as Arnold's test for *i.i.d.* models (Arnold et al., 1998), Ljung-Box for DTRW (Hamie et al., 2018), and tests described by Hoayek (2016). A third example, widely used in the literature, includes the ratio of high-to-low records which compares the frequency of higher records to lower records.

   (b) **Intensity-based features** refers to the examination of records beyond their mere count, with a focus on their magnitude and variability. For instance, the variance-to-mean ratio of records (VMR) which measures the variability of record values relative to their mean, with high values indicating significant jumps between records. Another feature is the simple dispersion measure of the whole time series X_t.

   (c) **Trend-based features** focus on the overall direction and progression of records. A trivial variable is the slope of the trend line to assess the overall direction of the time series computed as in Newman et al. (2010). Another feature is the normalized average of observed records (divided by the first trivial record) with high comparable values related to a steep increase in record values. Besides, the percentage increase in the values of records from the first to the last one within a time window also reflects the steepness of record increases regardless of intermediate fluctuations.

2. Features selection and weights

   An initial set of 19 features was developed based on theoretical and empirical considerations. For more information about the features, see Table B.1 in the appendix. These were refined through a two-stage process. First, features

exhibiting high multicollinearity or redundancy were identified and removed if pairwise correlation exceeds 0.9. This ensured that only independent and relevant variables were retained. Second, Exploratory Factor Analysis (EFA)[1] was performed on each dataset to identify the most statistically significant variables, to validate the grouping of features into the three primary factors (Frequency, Intensity, and Trend), and to assign weights (or factor loadings) to each feature. The analysis ensured that the 12 retained features explained more than 70% of the total variance. EFA was performed across multiple datasets, yielding slightly different sets of selected features and weights for each. To maintain consistency, the features most frequently selected across all datasets were retained for the final analysis (see Table B.2 in the appendix).

3. Scoring system

The final scoring system integrates the selected features into a uni-dimensional score, $\{S_t\}_{1\leq t\leq w}$, which reflects the combined influence of Frequency, Intensity, and Trend factors. For each factor, a weighted score is computed by summing the weighted values of its associated features and normalizing the result.

$$\text{Factor Score}_t = \sum(\text{Feature Value}_t \times w_i)/\sum w_i. \tag{1}$$

The overall final score is calculated as the average of the three normalized factor scores, ensuring balanced representation of Frequency, Intensity, and Trend.

$$S_t = \frac{1}{3} \times [\text{Frequency Score}_t + \text{Intensity Score}_t + \text{Trend Score}_t]. \tag{2}$$

To effectively capture temporal variations and address biases from initial trivial records, a moving window approach was adopted. For each sliding subset of the time series, a global score value was computed by summing the three factor scores. Hence, each point in the resulting score series S_t corresponds to a specific window, facilitating detailed analysis across geographical stations for both temperature and precipitation. The analysis began with a minimum window length of 30 observations (taking into consideration the asymptotic properties of records), sliding step-by-step through the series (the first window spanned the first 30 observations, the next started from the second observation to the 31st, and so on) and constructing the new score series. Sensitivity analysis using Dynamic Time Warping confirmed that varying window lengths had negligible effects on the resulting score series. For consistency, a window length of 50 observations was adopted, and windows with only one trivial record are ignored because of the impossibility to compute some features.

This methodology considers not only the timing, number, and intensity of climate records but also integrates non-record values, highlighting their significance in capturing fluctuations and underlying trends. The subsequent section will present the final model and analyze the scoring results for six geographical stations.

4 Data and Results

4.1 Data

The World Bank Climate Change Knowledge Portal provides complete yearly temperature and precipitation data from 1900 to 2021, including maximum, minimum, and average values. This analysis primarily focuses on average data, with maximum and minimum values specifically used to compute the "Ratio" feature[2]. To capture diverse climate patterns, six stations were selected to represent distinct Köppen-Geiger climate classifications[3]: Jambi, IDN (tropical), Cairo, EGY (arid), Beirut, LB (Mediterranean), Paris, FR (temperate), Oslo, NOR (cold) and Greenland (polar). Descriptive statistics for temperature and precipitation across these stations are summarized in Tables C.1 and C.2, while corresponding yearly trends are visualized in Figures C.1 and C.2.

Temperature levels exhibit an overall slightly increasing trend across most stations, a pattern supported by the KPSS test[4], which confirms non-stationarity (exceptions include Oslo and Greenland, where temperature trends appear stationary and sinusoidal, respectively). Temperature distributions were found to be Gaussian, as expected based on climatological literature. However, this assumption is not critical in our analysis, as our methodology does not impose any distributional constraints. On the other hand, precipitation levels show a stationary behavior across all locations —though they exhibit high short-term variability and localized fluctuations.

To develop the scoring model, the Egypt station dataset ($w = 72$ windows of T=50 observations) was used along with other datasets, but the former proved the best fit (higher proportion of variance explained, lower RMSEA and BIC; see Table B.2)for models comparison). EFA results identified three primary factors, which together explain over 80% of the total variance, validating our initial definition. The final model is as follow, before normalizing by each factor's weight and summing up as in eq. 1 and 2.

[1]Exploratory Factor Analysis (EFA) is a statistical technique used to identify underlying factors that explain the correlations among observed variables. It is a dimension reduction technique by grouping related features into common factors, making data interpretation more intuitive. Check Tucker and MacCallum (1997) for more info.

[2]The "Ratio" feature measures the proportion of high-to-low records in the dataset.

[3]The Köppen-Geiger classification system, developed by Wladimir Köppen, categorizes global climates into five major types.

[4]KPSS tests are used for testing a null hypothesis that an observable time series is stationary around a deterministic trend against the alternative of a unit root. For more info check Kwiatkowski et al. (1992)

$$\text{Frequency} = \mathbf{0.71}\ N_T + \mathbf{0.66}\ M + \mathbf{0.60}\ R,$$

$$\text{Intensity} = \mathbf{0.88}\ VMR_{X_t} + \mathbf{0.87}\ VMR_{X_t/R_n} + \mathbf{0.79}\ \overline{L_n},$$

$$\text{Trend} = \mathbf{0.82}\ \overline{R_n} + \mathbf{0.80}\ \alpha + \mathbf{0.92}\ P_{X_t,X_{L_n}} + \mathbf{0.79}\ P_{R_n,R_m} + \mathbf{0.90}\ VMR_{R_n} + \mathbf{0.84}\ \Delta X_t.$$

4.2 Climate Analysis Based on Generated Scores

By definition, scores are essentially linear combinations of various features, where high score values indicate an increased likelihood of encountering extreme weather events. Higher temperature scores indicate an increased frequency of unusually high temperatures, reinforcing evidence of global warming. Conversely, high precipitation scores differentiate between excessive rainfall (e.g., IDN, NOR, GRL) and prolonged dry periods (e.g., LB, EGY, FR), which might be problematic depending on the climate's natural characteristics.

Analyzing the temperature score series reveals a consistent global pattern across all stations, which was not readily apparent from raw temperature trends. Temperature scores exhibit an "inverted bathtub" shape, regardless of the specific depth or steepness at each station (see Figures D.1 and D.3 (a)). Generally, temperatures start the century with moderate to high scores, decline to a minimum around the mid-century period, and then gradually increase, often surpassing the initial values. The beginning scores align with the early warming phase of the 20th century. The dip in the mid-century may reflect the cooling period that occurred between 1940s and 1970s. During this time, global temperatures showed a decline or stagnation after the period of warming. The last increase corresponds to accelerated global warming, largely attributed to anthropogenic greenhouse gas emissions. A case study of Beirut (Figure 2) exemplifies this trend, showing a steady increase in temperature extremes post-1970s.

Conversely, precipitation scores exhibit diverse station-specific patterns (see Figures D.2 and D.3 (b))), reflecting regional climate dynamics. In Beirut (Figure 3), precipitation scores start low, then rise steadily before plunging in mid-century, coinciding with the lowest temperature scores. This suggests that cooler temperatures during this period may have enhanced condensation, leading to increased rainfall. Subsequently, the scores rise again, signaling a decline in precipitation and a shift toward drier conditions. Other stations exhibit different behaviors. In Indonesia and Greenland, precipitation scores show a consistently positive trend, indicating increasing rainfall intensity. This is likely driven by distinct climate processes; warming temperatures in Indonesia are making rainy seasons more intense while in Greenland, higher temperatures may be shifting precipitation from snow to rain. A different pattern emerges in Paris, where precipitation scores initially rise, indicating drier conditions, but then experience a sharp drop beyond the 1950s, signaling an increase in heavy rainfall events. This same pattern is observed in Oslo, though with greater intensity, suggesting a notable rise in extreme precipitation events in Northern Europe. In contrast, Cairo's precipitation scores have increased sharply in recent decades, mirroring a similar pattern observed in the mid-century. This suggests that Egypt's climate is becoming progressively drier.

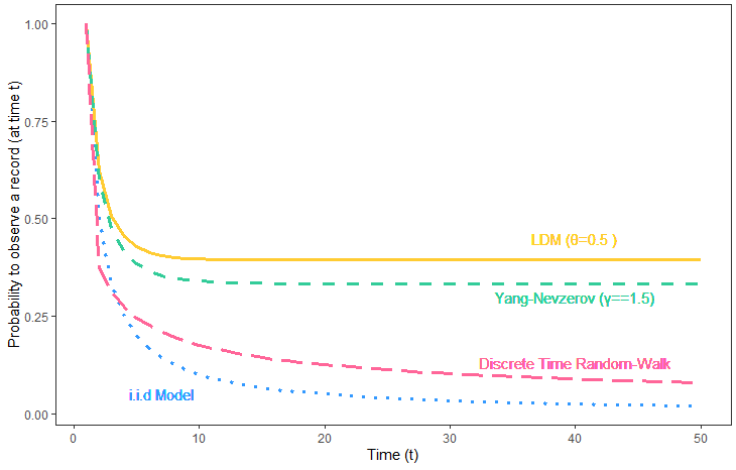
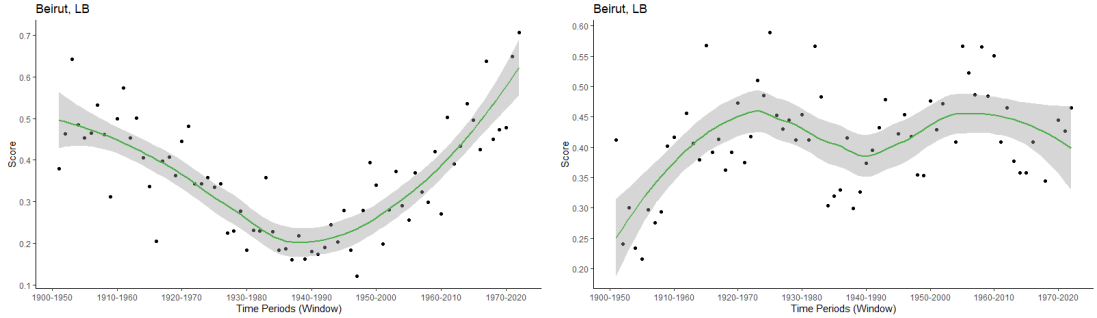


Figure 1: Probability of observing records at different time (t) acorss four different models



Overall, these patterns highlight significant climate changes across different regions. Table 1 summarizes the analysis about the scores obtained for each station. The variation in precipitation scores tells us how rainfall is changing according to different climate areas, while consistency in temperature scores reflect a common high temperature danger across the globe.

Table 1: Stations Scores Analysis

Station *(Climate)*	**Temperature Scores**	**Precipitation Scores**	**Interpretation**
Jambi, Indonesia *(Tropical)*	Stable and moderate initially, then decline mid-century, followed by a rapid increase.	Initially low, followed by a consistent increase.	Increased temperature and rainfall suggest intensifying extreme weather events, including heatwaves and flooding. This aligns with the expectation of a warmer, wetter climate in tropical regions.
Cairo, Egypt *(Arid)*	Initially high score and then rapid increase post deep dip mid-century.	Alternate between low initially, high in mid- century, and low again	The temperature increase after mid-century, coupled with precipitation fluctuations, indicates a growing risk of prolonged heatwaves and intensified drought periods. This pattern highlights increasing aridity and worsening water shortages.
Beirut, Lebanon *(Mediterranean)*	Start high, decline mid-century, and gradually increase, with a steeper rise in recent periods.	Initially low, fluctuate mid-century, then decline in recent decades.	Rising temperatures, along with decreasing precipitation in recent years, suggest a transition toward drier conditions, increasing the likelihood of droughts.
Paris, France *(Temperate)*	Initially low and stable, then rise sharply post-1950s.	Initially low, increase after the 1940s, then decline sharply in recent decades.	The post-1950s warming trend, combined with recent increased precipitation might indicate increased weather variability, potentially leading to more frequent heavy downpours.
Oslo, Norway *(Cold)*	U-shaped pattern, declining mid-century, then rising gradually at slow pace.	Fluctuate sinusoidally with a decline in recent periods.	Recent warming phase with increased rainfall may lead to heightened flood risks and changes in snowpack accumulation.
Greenland *(Polar)*	U-shaped curve, dipping mid-century with rapid increase afterwards.	Continuous increase since the beginning of the century.	Rapid temperature increases, alongside growing precipitation, signal accelerating ice melt. This contributes directly to rising sea levels. The shift in precipitation may also indicate a transition from snowfall to rainfall, further accelerating glacial retreat.

5 Conclusion

Our study employing records theory as a lens for examining climate change has provided valuable insights into the shifting patterns of temperature and precipitation across six diverse cities. Records theory, with its inherent properties and characteristics, has captured hidden climate behaviors that are not immediately apparent from traditional trend analysis. We have adapted three attributes to formulate a diverse set of variables, each designed to measure distinct aspects of weather patterns. Our scoring approach extended beyond only the frequency of records to their intensity, temporal patterns, and variability. Moreover, records were studied not only in isolation but also in relation to non-record events. The analysis of score levels underscores the urgency of addressing climate change on a global scale.

While our study has shed light on climate change, several promising avenues for future research in this area can enrich our understanding of the relationship between climate change and records theory. First, variables in the score we constructed can be refined. New variables can be added, and others can be removed or combined. Second, we can explore other methodologies than EFA or refine the weights assigned to variables within the current framework. The study was limited to six regions and two climate variables which can be extended to include more vulnerable regions and climate variables to provide broader insights. Multivariate framework can be built by mitigating the current framework and using a different definition of records, highlighting the interacting relationship between climate variables.

The score we developed is not exclusive to weather or climate data; we can explore its application to any dataset where changes in patterns and intensity need to be assessed. Replicating the study with a different dataset could yield interesting insights and validate the approach's applicability. Last but not least, while our study has demonstrated the efficacy of records theory in assessing climate change, collaborating with other statistical approaches, such as Extreme Value Theory (EVT), could unlock additional dimensions of insight and enhance our methodological framework. Integrating current climate models for simulations and forecasts with records theory could also improve the obtained model's results.

References

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (1998). *Records*. John Wiley & Sons.

Benestad, R. (2003). How often can we expect a record event? *Climate research*, 25(1):3–13.

Castillo-Mateo, J., Cebrián, A. C., and Asín, J. (2023). Statistical analysis of extreme and record-breaking daily maximum temperatures in peninsular spain during 1960–2021. *Atmospheric Research*, page 106934.

Chandler, K. (1952). The distribution and frequency of record values. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(2):220–228.

Elguindi, N., Rauscher, S., and Giorgi, F. (2013). Historical and future changes in maximum and minimum temperature records over europe. *Climatic Change*, 117:415–431.

Hamie, H., Hoayek, A., and Auer, H. (2018). Modeling the price dynamics of three different gas markets-records theory. *Energy strategy reviews*, 21:121–129.

Hoayek, A. (2016). *Estimation des paramètres pour des modèles adaptés aux séries de records*. PhD thesis, Université Montpellier.

IPCC (2023). *Summary for Policymakers*. IPCC, Geneva, Switzerland.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, 54(1-3):159–178.

Majumdar, S. N. and Ziff, R. M. (2008). Universal record statistics of random walks and lévy flights. *Physical review letters*, 101(5):050601.

Meehl, G. A., Tebaldi, C., Walton, G., Easterling, D., and McDaniel, L. (2009). Relative increase of record high maximum temperatures compared to record low minimum temperatures in the us. *Geophysical Research Letters*, 36(23).

Nevzorov, V. (1990). Records for nonidentically distributed random variables. In *Proceedings of the Fifth Vilnius Conference*, volume 2, pages 227–233.

Newman, W. I., Malamud, B. D., and Turcotte, D. L. (2010). Statistical properties of record-breaking temperatures. *Physical Review E*, 82(6):066111.

Redner, S. and Petersen, M. R. (2006). Role of global warming on the statistics of record-breaking temperatures. *Physical Review E*, 74(6):061114.

Schär, C., Vidale, P. L., Lüthi, D., Frei, C., Häberli, C., Liniger, M. A., and Appenzeller, C. (2004). The role of increasing temperature variability in european summer heatwaves. *Nature*, 427(6972):332–336.

Tucker, L. R. and MacCallum, R. C. (1997). Exploratory factor analysis. *Unpublished manuscript, Ohio State University, Columbus*, pages 1–459.

Wergen, G. and Krug, J. (2010). Record-breaking temperatures reveal a warming climate. *Europhysics Letters*, 92(3):30008.

Modélisation de la limite supérieure des températures à la surface de la Terre fondée sur la physique du climat et de leurs variations spatiales

Laurie Leterrier[1] & Marine Demangeot[2] & Nicolas Meyer[3] & Philippe Naveau[4] & Thomas Opitz[5]

[1] *BioSP, INRAE, France, laurie.leterrier@inrae.fr*
[2] *IMAG, Université de Montpellier, France, marine.demangeot@umontpellier.fr*
[3] *IMAG, Université de Montpellier, LEMON, Inria, France, nicolas.meyer@umontpellier.fr*
[4] *LSCE, IPSL, France, philippe.naveau@lsce.ipsl.fr*
[5] *BioSP, INRAE, France, thomas.opitz@inrae.fr*

Résumé

Le réchauffement climatique amplifie les événements météorologiques extrêmes, et la modélisation des pires scénarios atteignables sur Terre constitue un défi majeur pour la gestion des risques. Les données disponibles sur ces événements sont généralement très limitées. Dans ce travail, nous visons à estimer la distribution des températures les plus extrêmes, en nous concentrant sur la détermination de leur limite supérieure pour une localisation donnée, cette limite étant supposée finie en raison de lois physiques. Les approches standards basées sur les statistiques des valeurs extrêmes déterminent la borne supérieure à partir des trois paramètres d'une distribution des valeurs extrêmes ajustée aux maxima annuels de température. Cependant, les bornes supérieures estimées de cette manière sont souvent dépassées par de nouvelles données, même si les hypothèses du modèle sont valables. Cela ne permet donc pas de modéliser correctement les événements les plus extrêmes. Même la prise en compte de la variation spatiale des températures avec les coordonnées géographiques utilisées comme covariables ne résout pas le problème. En revanche, des travaux récents en physique du climat présentent une équation déterministe qui contraint la limite supérieure des températures en surface, quantité à relativement forte variation spatio-temporelle, par trois autres variables atmosphériques, plus lisses spatio-temporellement à grande échelle.

Nous reparamétrons les distributions de valeurs extrêmes pour que la borne supérieure apparaisse comme un de leurs trois paramètres. Après avoir calculé une limite supérieure à l'aide de la contrainte physique dans un premier temps, nous pouvons estimer de manière flexible les paramètres restants à l'aide d'une régression spatiale conditionnée par la borne

fixe dans une seconde étape. Nous mettons cela en œuvre à l'aide d'un modèle bayésien hiérarchique dans le cadre de l'approche INLA-SPDE.

Pour améliorer encore l'estimation et l'évaluation de l'incertitude, nous estimons tous les paramètres (y compris les bornes supérieures) conjointement dans un modèle statistique bayésien à contrainte physique. Nous modélisons le système des maxima annuels des températures en surface et de trois autres variables physiques à l'aide de distributions marginales de valeurs extrêmes reliées par une copule de valeurs extrêmes. Ce modèle déduit des configurations de paramètres cohérentes en combinant les connaissances physiques, les distributions a priori et les données, et peut être étendu pour permettre une spatialisation des paramètres modélisant les températures extrêmes, ces dernières étant géographiquement corrélées. Nous illustrons l'utilité de cette approche à l'aide de données de simulation d'un modèle climatique et de données de réanalyse.

Mots-clés. Théorie des valeurs extrêmes, températures extrêmes, statistique environnementale, statistique spatiale

Abstract

Global warming amplifies extreme weather events, and modeling worst-case scenarios is a major challenge for risk management. Available datasets of such events are usually very small. In this work, we aim to estimate the distribution of the most extreme temperatures, focusing on determining their upper bound at a given location, which is assumed to be finite due to physical constraints. Standard approaches based on Extreme-Value Statistics determine the upper bound from the three parameters of an extreme-value distribution fitted to annual temperature maxima. However, estimated upper bounds are often exceeded by new data, even if model assumptions are valid. This defeats useful assessment of worst-case events. Even allowing for spatial variation of temperatures with geographical coordinates as covariates does not solve the issue. By contrast, recent work in climate physics presents a deterministic equation constraining the upper bound of the spatiotemporally highly variable surface temperatures by spatiotemporally smoother large-scale atmospheric variables.

We reparametrize extreme-value distributions to let the upper bound be a parameter. After estimating an upper bound using the physical constraint in a first step, we can flexibly estimate the remaining parameters in a second step using spatial regression conditioned on the fixed bound. We implement this using a Bayesian hierarchical model in the INLA-SPDE framework.

To further improve estimation and uncertainty assessment, we estimate all parameters (including upper bounds) jointly in a physics-constrained Bayesian statistical model. We model the system of annual maxima of surface temperatures and other physical variables using marginal extreme-value distributions joined by an extreme-value copula. This model infers coherent parameter configurations by combining physical knowledge, prior distributions and data, and can be extended to allow for spatially varying parameters of geographically correlated temperature extremes. We illustrate its utility with climate model simulations

and reanalysis data.

Keywords. Modeling extremes, extremes temperatures, environmental statistics, spatial statistics

Texte long

1 Bornes supérieures des températures en surface

Le réchauffement climatique induit des événements climatiques extrêmes, dont l'intensité et la fréquence augmentent pour certains d'entre eux. Pour cela, il est nécessaire de pouvoir modéliser les pires scénarios atteignables sur Terre, mais les événements climatiques extrêmes sont par définition rares : peu de données sont donc disponibles, ce qui rend les études statistiques de ce type de phénomènes particulièrement difficiles. Il existe cependant des modèles statistiques plus adaptés à ce type de données, issus de la Théorie des valeurs extrêmes et détaillés dans le livre de Coles et al. (2001).

Dans ce travail, nous visons à estimer la distribution des températures les plus extrêmes, en nous concentrant sur la détermination de leur limite supérieure pour une localisation donnée, cette limite étant supposée finie en raison de lois physiques.

Nous illustrerons l'utilité et les limites des différentes modélisations de la loi statistique des extrêmes de la variable T_{max}, représentant la température maximale en surface, à l'aide de simulations issues d'un modèle climatique physique (IPSL-CM6A-LR) permettant de générer certaines variables physiques pour l'hémisphère nord et sur une période temporelle s'étendant de 1850 à 2059.

Des données de réanalyse (ERA5) sont également utilisées, *i.e.*, des sorties d'un modèle conditionné aux observations. Ces données sont disponibles de 1940 à 2024 et pour la Terre entière, avec une résolution spatiale plus fine que celle du jeu de données précédent.

1.1 Borne statistique par maximum de vraisemblance

Les approches standards basées sur les statistiques des valeurs extrêmes déterminent la borne supérieure à partir des trois paramètres d'une distribution des valeurs extrêmes ajustée aux maxima annuels de température. En notant T_j la température journalière maximale de l'année j, la distribution de T_j peut être approchée par une loi GEV (Generalized Extreme Value) de paramètre de localisation μ, d'échelle σ et de forme ξ. La fonction de répartition d'une telle loi est donnée par :

$$G(t) = \mathbb{P}(T_j \leq t) = \exp(-[1 + \xi(\frac{t-\mu}{\sigma})]^{-1/\xi}), \tag{1}$$

avec t tel que $1 + \xi(\frac{t-\mu}{\sigma}) > 0$, $\mu \in]-\infty; +\infty[$, $\sigma > 0$ et $\xi \in]-\infty; +\infty[$.

Le paramètre de forme ξ va déterminer la forme de la queue de distribution de la variable aléatoire. Si ξ est strictement négatif, la queue de distribution est dite courte et la loi étudiée appartient à la famille de Weibull. Le support de la variable aléatoire est alors borné à droite par : $B_{stat} = \mu - \frac{\sigma}{\xi}$. Cette quantité correspond à la borne supérieure de T_j calculée d'un point de vue statistique, et implique $\mathbb{P}(T_j > B_{stat}) = 0$. C'est donc cette borne que l'on souhaite déterminer, en estimant les trois paramètres (μ, σ, ξ) de la GEV grâce aux mesures des températures maximales annuelles passées.

En estimant les bornes supérieures par maximum de vraisemblance et de façon stationnaire, nous remarquons que ces estimations sont finalement dépassées par les températures futures. Cet événement étant censé se produire avec une probabilité nulle, il illustre les limites du modèle utilisé et la mauvaise modélisation des événements les plus extrêmes. Même en essayant d'améliorer cette modélisation par l'ajout de covariables physiques ou par la prise en compte de la spatialité des températures, ce problème persiste.

La figure 1 illustre le dépassement de la borne prédite. Pour chaque pixel, les trois paramètres de la loi GEV, et donc la borne supérieure, sont estimés de manière stationnaire (sans covariable) à partir des températures maximales annuelles comprises entre 1945 et 2014. Une localisation géographique sera représentée en rouge si la borne estimée s'avère être dépassée au moins une fois sur la période temporelle hors période d'apprentissage, donc entre 1850 et 1944 et entre 2015 et 2059. Elle sera représentée en noir sinon.

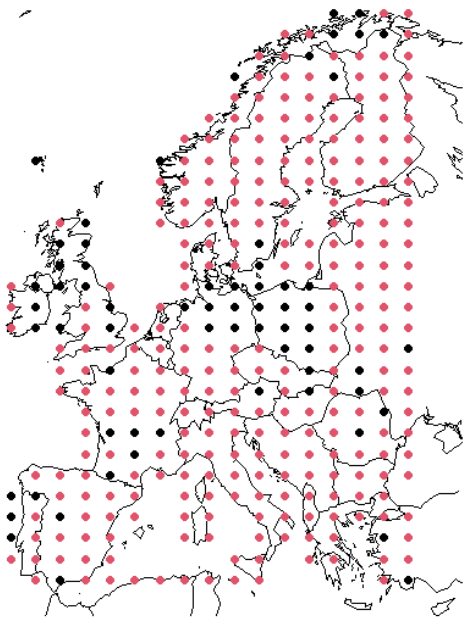


Figure 1: Carte de l'Europe, où les points rouges correspondent à un dépassement de la borne prédite

La figure 1 montre que la majorité des bornes prédites en Europe, par maximum de vraisemblance et de façon stationnaire, sont dépassées (points rouges) au moins une fois, ce qui incite à proposer une autre méthode pour estimer cette borne supérieure.

1.2 Borne physique

Des travaux récents en physique du climat, présentés dans l'article de Zhang et William (2023) et dans celui de Noyelle et al. (2024), introduisent une équation déterministe qui contraint la limite supérieure des températures en surface, quantité à forte variation spatio-temporelle, par des variables atmosphériques, plus lisses spatio-temporellement à grande échelle. L'équation proposée est la suivante :

$$T_{max} = T_{500} + \frac{L_v}{c_p}(Q_{sat}(T_{500}) - Q) + \frac{g}{c_p}(Z_{500} - Z_s), \tag{2}$$

avec T_{max} la température maximale en surface, T_{500} la température de l'air à une pression de 500 hPa, $Q_{sat}(T_{500})$ l'humidité spécifique de saturation à 500 hPa, Q l'humidité spécifique en surface de la parcelle d'air, Z_{500} la hauteur du géopotentiel à 500 hPa, Z_s l'élévation de la surface, L_v la constante de chaleur latente de vaporisation, c_p la constante de chaleur spécifique de l'air à pression constante, et g la constante d'accélération gravitationnelle.

De ce fait, une manière de trouver le majorant de la température maximale pouvant être atteinte est de maximiser l'équation (2), en maximisant T_{500} et Z_{500} et en minimisant Q, toutes les constantes de l'équation physique étant positives. Cependant, le majorant calculé ainsi n'est pas forcément la borne supérieure car la combinaison ($T_{500_{max}}$, $Z_{500_{max}}$, Q_{min}) n'est peut-être pas atteignable physiquement. On aura alors tendance à surestimer la borne maximale des températures annuelles, mais ce choix est préférable à une sous-estimation, qui pourrait provoquer des conséquences plus importantes.

2 Méthodes utilisées

2.1 Reparamétrisation du problème

Nous reparamétrons les distributions de valeurs extrêmes pour que la borne supérieure soit un paramètre. Après avoir estimé une limite supérieure à l'aide de la contrainte physique (2) dans un premier temps, nous pouvons estimer les paramètres restants à l'aide d'une régression spatiale conditionnée par la borne fixe dans une seconde étape.

En effet, au lieu d'ajuster une loi GEV aux données composées des températures maximales annuelles, nous pouvons chercher à modéliser par une loi de Weibull les quantités $\tilde{T}_j = B - T_j$, *i.e.*, l'écart des températures à la borne fixe B estimée précédemment. De ce fait, nous cherchons à estimer les paramètres de forme et d'échelle d'une loi de Weibull, ces derniers nous permettant alors d'accéder à une estimation des paramètres σ et ξ de la loi GEV utilisée initialement. Le problème ne revient donc plus à estimer (μ, σ, ξ), mais à le reparamétrer par (B, σ, ξ). En utilisant des connaissances sur B, ici en la calculant physiquement, il ne reste que deux paramètres à estimer.

Nous mettons cela en œuvre à l'aide d'un modèle hiérarchique bayésien dans le cadre INLA-SPDE, méthode expliquée dans Huang et al. (2017). Avec ce modèle, les estimations

ne sont pas calculées de façon indépendante pixel par pixel, permettant ainsi de tenir compte de la cohérence spatiale des températures.

La figure 2 est obtenue après avoir appliqué la méthode INLA à la variable $\tilde{T}_j$. Par son implémentation sur le logiciel R, cette méthode permet de faire dépendre de covariables le paramètre d'échelle de la loi de Weibull, mais le paramètre de forme de cette même loi est quant à lui estimé de manière constante.

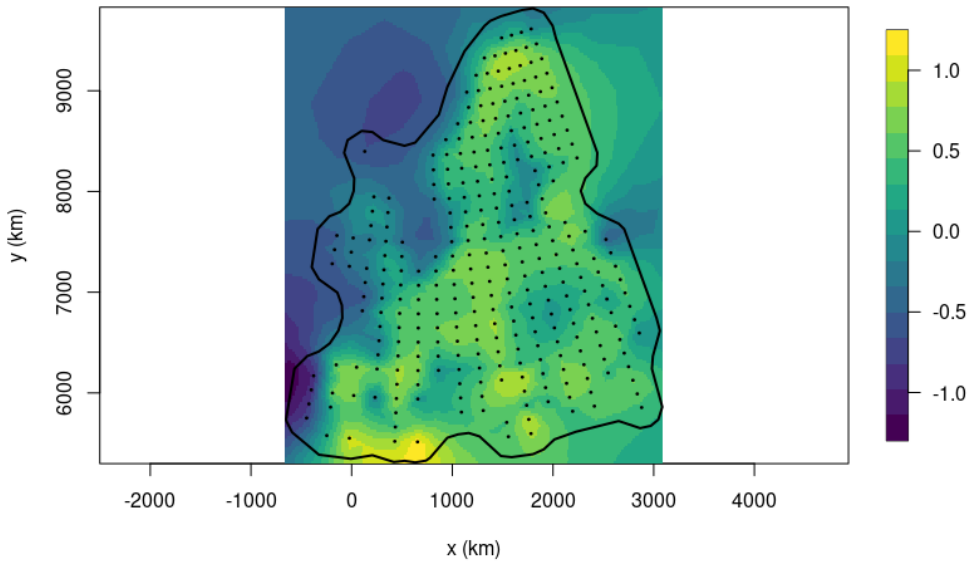


Figure 2: Effet spatial obtenu avec la méthode INLA

Il semble y avoir une augmentation de l'effet spatial lorsque l'on va du nord vers le sud, bien qu'elle ne soit pas élevée. Cela est cohérent avec les connaissances disponibles, les températures ont tendance à changer en fonction de la latitude, mais pas nécessairement au niveau de la longitude.

2.2 Modélisation des covariables physiques et de leur dépendance

La théorie des valeurs extrêmes multivariées, détaillée dans le livre de Beirlant et al. (2006), permet de modéliser la distribution jointe de variables prenant des valeurs extrêmes ainsi que leur dépendance. Pour améliorer encore l'estimation et l'évaluation de l'incertitude, nous estimons tous les paramètres (y compris les bornes supérieures) conjointement dans un modèle statistique bayésien à contrainte physique. Nous modélisons le système des maxima annuels des températures en surface et d'autres variables physiques, présentes dans l'équation (2), à l'aide de distributions marginales de valeurs extrêmes reliées par une copule de valeurs extrêmes. En particulier, chaque variable sera représentée par ses maxima (ou minima pour Q) annuels, modélisés par une loi GEV et possédant donc trois paramètres à estimer. De plus, des paramètres supplémentaires devront être estimés pour modéliser la dépendance entre les variables à travers la copule.

Ce modèle déduit des configurations de paramètres cohérentes en combinant les connaissances physiques, les distributions a priori sur chaque paramètre et les données, et peut être étendu pour permettre une spatialisation des paramètres modélisant les températures extrêmes, ces dernières étant géographiquement corrélées.

Dans les résultats présentés, nous comparerons les différentes approches de modélisation statistique en utilisant les données mentionnées plus haut, et nous analyserons leur capacité à estimer une borne supérieure plausible pour les températures maximales en surface.

Bibliographie

Beirlant, Jan and al. (2006), Statistics of extremes: theory and applications, *John Wiley & Sons*

Coles, Stuart and al. (2001), An introduction to statistical modeling of extreme values, *Springer*, Vol. 208.

Huang, Jingyi et al. (2017), Evaluating a Bayesian modelling approach (INLA-SPDE) for environmental mapping, *Science of The Total Environment*

Noyelle, Robin and al. (2024), Integration of physical bound constraints to alleviate shortcomings of statistical models for extreme temperatures, *hal-04479249*

Zhang, Yi and William, R Boos (2023), An upper bound for extreme temperatures over midlatitude land, *Proceedings of the National Academy of Sciences*

Extrêmes et environnement

CONDITIONS ENVIRONNEMENTALES EXTRÊMES DANS LE RAZ BLANCHARD

Paul-Adrien Alves[1] & Nicolas Raillard[2]
[1]LHyMar, Ifremer, France, paul.adrien.alves@ifremer.fr
[2]LHyMar, Ifremer, France, nicolas.raillard@ifremer.fr

Résumé

Le Raz Blanchard, un détroit situé dans la Manche entre la péninsule du Cotentin (Normandie) et l'île anglo-normande d'Aurigny, est réputé pour ses forts courants de marée (jusqu'à 5 m/s) et son potentiel énergétique significatif [1, 2]. S'ajoutent à cela des états de mer complexes et la survenue occasionnelle d'événements extrêmes, à l'image de la tempête Ciarán (2023), qui représentent des risques majeurs aussi bien pour les structures offshore que les structures sous-marines.

Cette étude, menée dans le cadre d'une thèse, présente les premiers résultats de la caractérisation des événements de vagues extrêmes et de leurs interactions avec les courants de marée [3, 4] en utilisant la base de données RESOURCECODE [5], qui regroupe 27 ans de données issues de modèles numériques à haute résolution. Des lois de Pareto généralisées combinées aux modèles additifs généralisés sont utilisés pour analyser les hauteurs significatives des vagues.

Les résultats montrent que les vagues extrêmes coïncident souvent avec les courants les plus rapides, de nombreux événements se produisant lorsque la direction des vagues est opposée à celle du courant. Ces conditions amplifient les risques pour les structures posées sur le fond marin, soulignant la nécessité de prendre en compte les valeurs de retour des événements extrêmes dans la conception des structures pour garantir leur durabilité à long terme.

Abstract

The Alderney Race, a strait located in the English Channel between the Cotentin Peninsula (Normandy) and the Channel Island of Alderney, is renowned for its strong tidal currents (up to 5 m/s) and significant energy potential [1, 2]. Not only that, this region also experiences rough seas and occasionally extreme events, such as Storm Ciarán (2023), which pose major risks to both offshore and underwater structures.

This study, as part of a doctoral thesis, consists in the first results in characterizing extreme wave events and their interactions with tidal currents [3, 4] using the RESOURCECODE database [5], which compiles 27 years of high-resolution numerical model data. Generalized Pareto distributions combined with generalized additive models are used to analyze significant wave heights.

Results reveal that extreme waves often coincide with the fastest currents, with many events occurring when wave and current directions oppose. These conditions amplify risks to seabed

structures, highlighting the need to consider return values of extreme events in structural design for long-term durability.

Mots-clés

Valeurs extrêmes, Océanographie, Hauteur significative des vagues, GAM, Valeur de retour, R.

Keywords

Extreme values, Oceanography, Significant wave height, GAM, Return values, R.

1 Introduction

Dans le cadre d'un laboratoire commun avec la société HydroQuest, entreprise de développement de structures pour l'hydrolien marin, l'Ifremer effectue des travaux d'observation de suivis (réglementation sur l'accès, impact sur la biodiversité, relevés de caractéristiques physico-chimiques de l'eau...) réalisés sur différentes fermes hydroliennes, de la modélisation numérique des structures avec de l'expérimentation en bassin pour validation, ainsi que des analyses statistiques ayant pour objectif de caractériser les conditions environnementales du Raz Blanchard, lieu de passage de courants parmi les plus rapides au monde situé dans la Manche, entre la péninsule du Cotentin (Normandie) et l'île anglo-normande d'Aurigny, dont le potentiel énergétique pour l'énergie hydrolienne est bien identifié [1]. Il s'agit aussi bien de décrire les états de mer que de calculer les probabilités de retour d'événements extrêmes, en raison des contraintes mécaniques qu'ils sont suspectibles d'exercer sur les hydroliennes qui seront posées.

2 Base de données

Les analyses s'appuient toutes sur la base de données Resourcecode [5]. Celle-ci se constitue de rejeux d'états de mer (i.e. prévisions numériques sur des situations passées avec un modèle météorologique). Les données de RESOURCECODE s'appuient sur les modèles numériques WAVEWATCH-III (vagues), ERA5 (vent), ainsi que MARS2D (courants de marée). La base de données Resourcecode dispose d'une très haute résolution, permettant de sélectionner précisément le point adéquat. Le modèle réalise une prévision toutes les heures sur toute la période 1994-2023 (nous ne disposons que de la période 01/01/1994 - 31/12/2020). Nous n'utilisons qu'un seul de ces points dans la présente analyse.

Les principales variables que nous utilisons sont des paramètres d'états de mer, calculés à partir de spectres d'états de mer générés par le modèle WAVEWATCH-III. La variable dont nous chercherons à estimer les extrêmes est la hauteur significative des vagues H_s, définie empiriquement comme la moyenne du tiers des vagues les plus hautes. Nous nous intéresserons à son lien avec la vitesse du courant C_{spd} et $EAVC$, l'écart angulaire vagues-courant. La distribution des H_s est disponible en Figure 1.

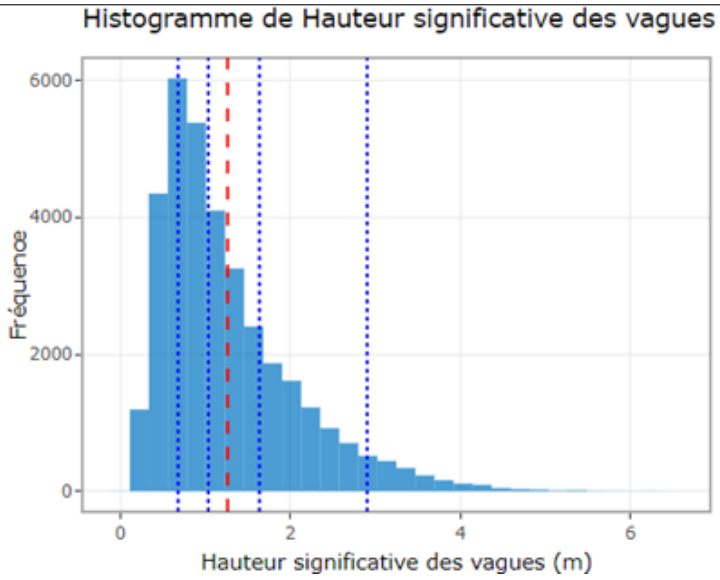


FIGURE 1 – Distribution des H_s sur la période 1994-2020 (données horaires). Les tirets rouges indiquent la moyenne, tandis que les pointillés bleus représentent les quantiles 0,25, 0,5, 0,75 et 0,99.

3 Ajustement multivarié des excès au seuil

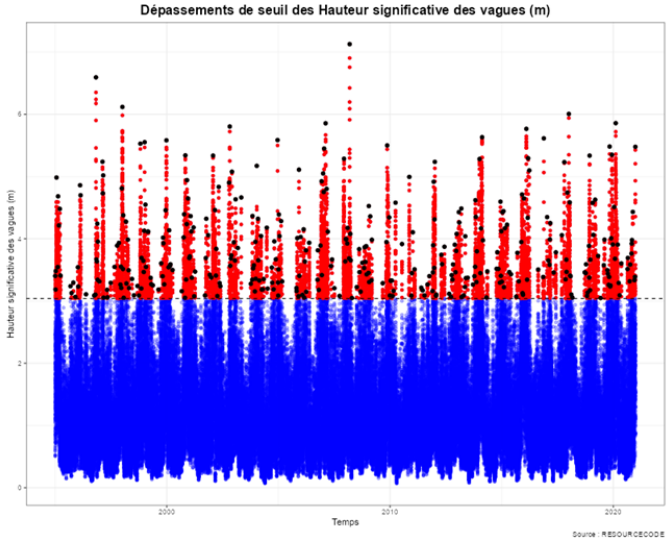


FIGURE 2 – Dépassements de seuil des H_s sur la période 1994-2020 (données horaires). Les points rouges indiquent les dépassements de seuil et les points noirs sont les maxima des clusters identifiés selon la méthode *runs* [6].

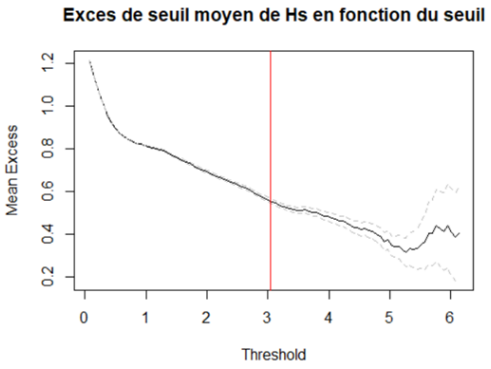


FIGURE 3 – Excès de seuil moyen en fonction du seuil. En rouge, le seuil choisi, $u = 3,04$.

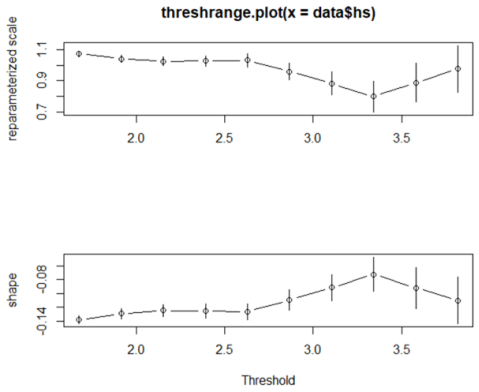


FIGURE 4 – Évolution des erreurs-type autour des paramètres selon le seuil.

Pour la modélisation des H_s extrêmes, nous tirerons parti de la flexibilité des modèles additifs généralisés, dont la forme classique est la suivante [7] :

$$g(\mu_i) = \boldsymbol{X}\beta + \sum_{j=1}^{n} f_j(x_{j,i})$$

où g une fonction de lien, $\mu_i = \mathbb{E}(Y_i|x_i)$ et f_j est représentée par m fonctions de base pondérées, avec $b_{j,k}$ la k-ème fonction de base, pondérée par $\beta_{j,k}$, de sorte que $f_j(x_j) = \sum_{k=1}^{m} b_{j,k}(x_j)\beta_{j,k}$.

La vraisemblance du modèle devient

$$l(\lambda) = l(\hat{\beta}_\lambda, \lambda) + \frac{1}{2}\log|S_\lambda|_+ - \frac{1}{2}\log|H(\hat{\beta}_\lambda)| + C, \tag{1}$$

où $\hat{\beta}_\lambda$ maximise la vraisemblance pénalisée $l(\beta_\lambda, \lambda) = l(\beta_\lambda) - \frac{1}{2}\beta^T S_\lambda \beta$ pour un λ donné, avec $S_\lambda = \sum_{i=1}^{k} \lambda_i S_i$ matrice de pénalisation pour les fonctions de base, $|S_\lambda|_+$ le produit des valeurs propres positives de S_λ, $H(\hat{\beta}_\lambda) = -\nabla\nabla^T l(\beta, \lambda)|_{\beta=\hat{\beta}_\lambda}$ le produit des gradients $\nabla\nabla^T$ de la vraisemblance évalué en $\hat{\beta}_\lambda$ et C est une constante. Le paramètre de lissage optimal $\hat{\lambda}$ est obtenu par optimisation numérique via la méthode de quasi-Newton BFGS.

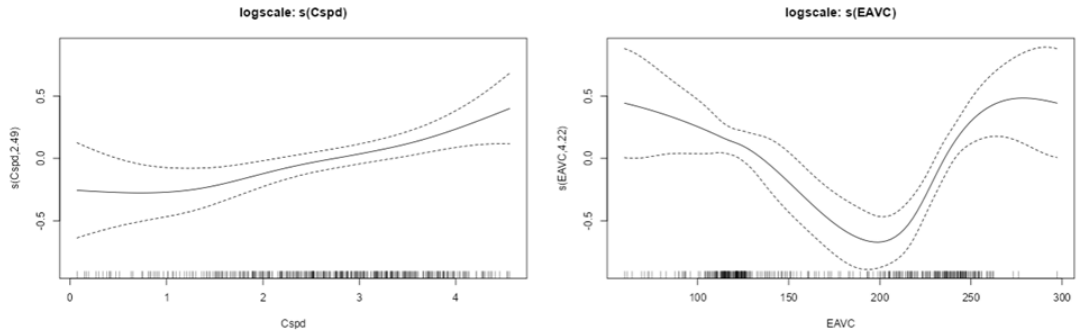


FIGURE 5 – Évolution de la valeur des paramètres estimés par modèle additif généralisé de la loi GP selon les covariables.

Nous constatons ci-dessus que le paramètre d'échelle augmente avec les vitesses de courant élevées. La variance d'estimation est importante pour les valeurs de vitesses de courant faibles,

pour lesquelles nous avons peu de maxima de clusters (voir Figure 6). Le paramètre d'échelle monte également pour des valeurs d'écart angulaire autour de 120 et de 250 et est minimal autour de 200 (opposition vagues-courant), cependant la variabilité d'estimation est importante lorsque les vagues et courant vont dans le même sens (EAVC=0/360°) ou sont en opposition (EAVC=180°) car au Raz Blanchard, le courant est essentiellement perpendiculaire aux vagues. Il y a donc peu de maxima de clusters en ces lieux, comme en atteste la Figure 6.

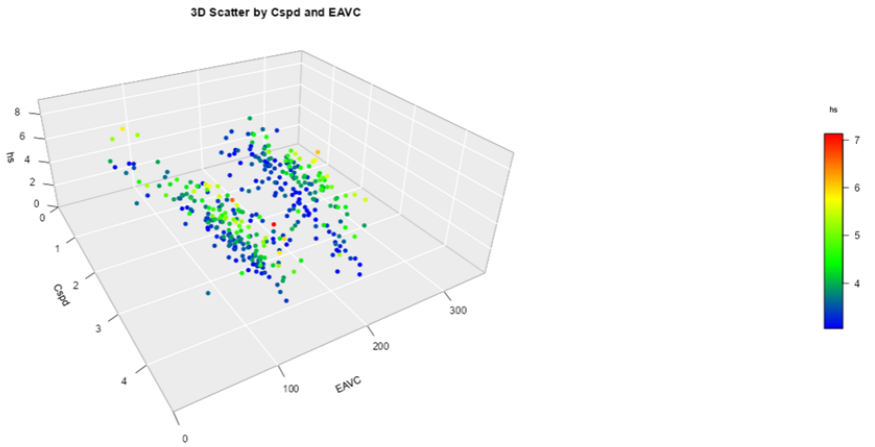


FIGURE 6 – Maximum des clusters de H_s conservés pour l'inférence.

La Figure 7 présente les valeurs de retour toutes les 8760 observations (nous disposons de 24 données par jour fois 365 jours) des lois de Pareto généralisées ajustées aux H_s, dont les paramètres ont été estimés pour des combinaisons des covariables $(C_{spd}, EAVC)$ proches de celles observées dans les maxima de clusters afin d'éviter des calculs où la variabilité est trop importante.

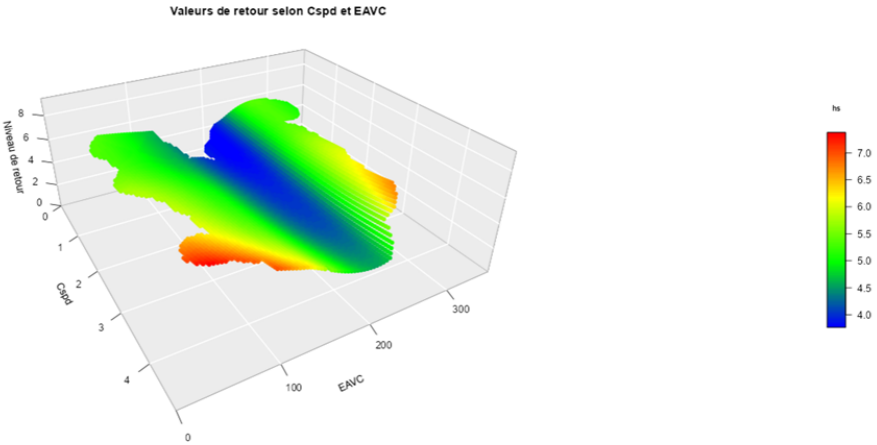


FIGURE 7 – Niveau de retour toutes les 8760 observations des lois de Pareto généralisées ajustées aux H_s dépendant des combinaisons de $(C_{spd}, EAVC)$ non trop distantes des maxima de clusters.

Ce modèle permet d'identifier les effets de la vitesse du courant et de l'écart angulaire vagues-courant sur les H_s engendrées. L'interprétation ponctuelle est en revanche difficile puisque

nous avons estimé des paramètres d'échelle de lois de Pareto généralisées différents selon chaque combinaison de covariables et calculé le retour à 1 an en partant du principe que toute l'année se déroulait sous les mêmes combinaisons de covariables, ce qui n'est pas le cas en pratique, i.e. nous avons calculé le niveau de retour [8], avec $m = 365,25 \cdot 24 = 8760$, tel que

$$\hat{z}(C_{spd}, EAVC) = \frac{\hat{\sigma}(C_{spd}, EAVC)}{\hat{\xi}}((m\theta\zeta_u)^{\hat{\xi}} - 1)$$

avec θ l'indice extrémal et ζ_u la proportion des dépassements de seuil. Une amélioration notable, qui constitue la prochaine étape de ce travail, consiste à passer à une approche qui intégrerait la fréquence de dépassement de seuil selon chacune des combinaisons d'états de mer, et à calculer des valeurs de retour interprétables (i.e. si l'on estime qu'une combinaison d'états de mer est observée en moyenne 20 fois dans l'année, alors le retour à 1 an correspondra à la valeur dépassée en moyenne une seule fois sur les 20 chaque année).

4 Bibliographie

Références

[1] Myers, L., & Bahaj, A. S. (2005). Simulated electrical power potential harnessed by marine current turbine arrays in the Alderney Race. *Renewable Energy*, 30(11), 1713–1731.

[2] Furgerot, L., et al. (2020). One year of measurements in Alderney Race. *Philosophical Transactions : Mathematical, Physical and Engineering Sciences*, 378(2178), 1–20.

[3] Mackay, E., & Hardwick, J. (2022). Joint Extremes of Waves and Currents at Tidal Energy Sites in the English Channel. *OMAE2022-79348*.

[4] Bennis, A. C., et al. (2020). Numerical modelling of three-dimensional wave-current interactions in complex environments : application to Alderney Race. *Applied Ocean Research*, 95, 102021.

[5] Accensi, M., et al. (2021). ResourceCODE framework : A high-resolution wave parameter dataset for the European Shelf and analysis toolbox. In *14th European Wave and Tidal Energy Conference (EWTEC)*, pp. 2182-1.

[6] Ferro, C. A., & Segers, J. (2003). Inference for clusters of extreme values. Journal of the Royal Statistical Society Series B : Statistical Methodology, 65(2), 545-556.

[7] Wood, S. N. (2017). Generalized additive models : an introduction with R. chapman and hall/CRC.

[8] Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. London : Springer.

Extended extreme-value models with spatial dependence

Lorenzo Dell'Oro [1], Carlo Gaetan [2] & Thomas Opitz [3]

[1] *Università di Padova, Italie, lorenzo.delloro@phd.unipd.it*
[2] *Università Ca' Foscari di Venezia, Italie, gaetan@unive.it*
[3] *INRAE BioSP, France, thomas.opitz@inrae.fr*

Résumé. La modélisation spatiale des valeurs extrêmes permet d'étudier le risque d'occurrence conjointe d'événements extrêmes dans différents sites et présente un intérêt significatif pour les sciences climatiques et les autres sciences de l'environnement. Une classe populaire de modèles, en raison de sa flexibilité et de son importance pour les représentations asymptotiques, est celle des mélanges d'échelle ou de position, dans lesquels un champ aléatoire spatial de référence est multiplié ou translaté, respetivement, par une variable aléatoire, ce qui peut modifier son comportement de dépendance extrémale. Les mélanges gaussiens d'échelle et position (*location-scale mixtures* en anglais) conservent beaucoup des avantages de leur processus gaussiens de référence, tout en surmontant certaines de leurs limites, telles que la symétrie, les queues de distribution légères et la faible dépendance extrémale. Dans ce travail, nous passons en revue les propriétés des mélanges gaussiens d'échelle et position et développons plusieurs nouvelles constructions présentant des caractéristiques intéressantes. Nous tirons parti de leur flexibilité pour proposer des *modèles de valeurs extrêmes étendus*, qui permettent de modéliser de manière appropriée non seulement les queues de distribution, mais aussi le corps de la distribution des données. Cela permet d'éviter l'incertitude liée à la nécessité de sélectionner un seuil artificiel pour définir les valeurs extrêmes en les séparant des évévnements ordinaires dans le corps de distribution. Nous proposons également de nouvelles solutions pour l'inférence de la vraisemblance dans les modèles paramétriques de mélanges gaussiens d'échelle et de position, afin d'éviter le goulot d'étranglement numérique donné par les variables latentes de d'échelle et de position qui conduisent à un coût de calcul élevé des évaluations de la vraisemblance standard. L'efficacité des modèles et des méthodes d'inférence est confirmée par des exemples de données simulées, et nous présentons une application aux variables météorologiques liées aux incendies de forêt.

Mots-clés. Extrêmes spatiaux, Processus gaussiens, Dépendance et indépendance asymptotiques

Abstract. Spatial modeling of extreme values allows studying the risk of joint occurrence of extreme events at different locations and is of significant interest in climatic and other environmental sciences. A popular class of models, due to its flexibility and importance for asymptotic representations, is given by random location-scale mixtures of Gaussian processes. The general principle of a location or scale mixture is that a spatial "baseline" process is multiplied or shifted by a random variable, thus modifying its marginal distributions and its extremal dependence behavior. Gaussian location-scale mixtures retain benefits of their Gaussian baseline processes while overcoming some of their limitations, such as symmetry,

light tails and weak tail dependence. We first review properties of Gaussian location-scale mixtures, and then develop several novel constructions with interesting features. We leverage their flexibility to propose extended extreme-value models that allow for appropriately modeling not only the tails but also the bulk of the data. This avoids the uncertainty brought by the need to select an artificial threshold to define extreme events and separate the bulk from the tail. We also propose new solutions for likelihood inference in parametric models of Gaussian location-scale mixtures, in order to bypass the numerical bottleneck given by the latent location and scale variables that lead to high computational cost of standard likelihood evaluations. The effectiveness of the models and of the inference methods is confirmed with simulated data examples, and we present an application to wildfire-related weather variables.

Keywords. Spatial extremes, Gaussian processes, Asymptotic dependence and independence

1 Modeling of spatial extremes

When studying extreme values in random fields, the strength of the spatial dependence can be characterized by the (bivariate) χ-measure or *coefficient of upper tail dependence* (Huser and Wadsworth, 2022). Let $Y(s)$ be a stochastic process defined on a nonempty domain $\mathcal{S}$ with $s_1, s_2 \in \mathcal{S}$; then,

$$\chi(s_1, s_2) = \lim_{u \to 1} \chi_u(s_1, s_2) = \lim_{u \to 1} \Pr\left(Y(s_1) > F_1^{-1}(u) | Y(s_2) > F_2^{-1}(u)\right),$$

where $Y(s_1) \sim F_1$, $Y(s_2) \sim F_2$, such that $F_1(Y(s_1)), F_2(Y(s_2)) \sim \text{Unif}(0,1)$ and $u \in (0,1)$. If $\chi(s_1, s_2) > 0$, $Y(s_1)$ and $Y(s_2)$ are *asymptotically dependent*, while if $\chi(s_1, s_2) = 0$, they are *asymptotically independent*. In the first case, the spatial dependence strength of the process does not vanish as events become more extreme, which may be not realistic in many environmental application. This has led to the development of *sub-asymptotic* models, that try to accurately describe not only the limiting dependence, but also the joint tail decay rate of spatial processes (Huser and Wadsworth, 2022).

The most widely used class of spatial processes are Gaussian processes. They always display asymptotic independence, due to the light joint tail of the multivariate Gaussian distribution, unless they are characterized by perfect dependence (i.e., a correlation function given as $\rho(s_1, s_2) = 1$). Therefore, the limiting coefficient $\chi(s_1, s_2)$ for non-trivial Gaussian processes is always equal to 0, even though its decay rate in the subasymptotic coefficient $\chi_u(s_1, s_2)$ is controlled by the correlation function $\rho(s_1, s_2)$.

A flexible and widely used family of sub-asymptotic models is the one of *random scale mixtures* $X(s) = RW(s)$, $s \in \mathcal{S}$, in which a generic spatial process $W(s)$ is multiplied by an independent positive random variable R. Their extremal dependence properties have been studied in the bivariate (non-spatial) case by Engelke et al. (2019). An interesting case is that of Gaussian scale mixtures (Huser et al., 2017), in which $W(s)$ is a standard Gaussian process, but the tail heaviness of the random variable R (which is common over

all the spatial domain) increases the strength of extremal dependence and induce asymptotic dependence in the process $X(s)$ for certain choices of the scale variable. An alternative class is that of Gaussian location mixtures (Krupskii et al., 2018), $X(s) = S + W(s),\ s \in \mathcal{S}$, with $W(s)$ a Gaussian process and S a random variable independent of $W(s)$. Up to a marginal logarithmic transformation, a random location mixture is equal to a random scale mixture with scale $\exp(S)$ and process $\exp(W(s))$.

2 Gaussian location-scale mixtures

We can define general Gaussian location-scale mixtures as

$$X(s) = S + R \times W(s), \quad s \in \mathcal{S}, \tag{1}$$

where $\boldsymbol{W} = \{W(s)\} \sim \mathcal{GP}$ and $R \geq 0$, and W is independent of R, S. Usually, we also assume independence of R and S, although some model classes do not make this assumption, for example so-called normal mean-variance mixtures where $S = cR^2$ with $c \in \mathbb{R}$.

Models of type (1) are not limited by the symmetry, light tails and weak tail dependence of Gaussian processes, because the random variable S and R can alter the tail behavior of $W(s)$, possibly in different ways for the two tails. Nevertheless, these more flexible models retain benefits of Gaussian processes, such as a parametrization depending on a covariance function, easiness of conditional and unconditional simulation, and a relativey straightforward adaptation of already existing implementations of modeling and prediction tools. However, the presence of the latent variables S and R can lead to a high numerical cost of likelihood evaluations if no closed-form densities are available, which is the case with a few well-known exceptions.

Let G be the marginal distribution $X(s)$ in (1). Then we can define the model we fit to data on a marginally transformed scale as follows:

$$Y(s) = F_s^{-1}(G(X(s))), \tag{2}$$

for any generic marginal distribution F_s of the data process $Y(s)$, i.e., we apply a copula approach. Therefore, $Y(s)$ is on the scale on which data are observed, while the purpose of the model (1) is to describe the dependence structure of the data, irrespective of their marginal scale. An interesting example of marginal distribution F_s is the *Extended Generalized Pareto Distribution* (Naveau et al., 2016), which models the complete distribution of the data, without the need to select an artificial threshold to define extreme values and separate the bulk from the tail.

Example (Gaussian scale mixture) Let $E \sim \text{Exp}(1/2)$, $G \sim \text{Gamma}(1/\xi, 1/\xi)$, and $R = \sqrt{E}/G$. Then the scale mixture $X(s) = RW(s)$ follows marginally a Symmetric (or two-sided) Generalized Pareto Distribution, $\text{SGPD}(\xi, 1)$, and it is asymptotically dependent, despite the asymptotic independence of the Gaussian process $W(s)$. Moreover, as $\xi \to 0$, G converges to 1 and $X(s)$ tends to a Laplace random field, which is asymptotically independent. Therefore, there is a transition between the two extremal dependence classes, although it occurs at the boundary of the parametric space.

3 New solutions for inference

Let the set of observed locations be $\mathcal{S} = \{s_0, \ldots, s_m\}$. If we first estimate the marginal distribution $\hat{F}_s$ of $Y(s)$ in (2), we can reconstruct the process $X(s)$ as

$$X(s) = G^{-1}\left(\hat{F}_s(Y(s))\right),$$

noting that, for fixed values of the parameters of the random variables S and R in (1), G is known, being $W(s)$ marginally distributed as a standard Gaussian.

3.1 Inference for the parameters of $W(s)$

We can distinguish three cases: Gaussian location mixtures (when $R = 1$), Gaussian scale mixtures (when $S = 0$) and general Gaussian location-scale mixtures.

- If $R = 1$, model (1) becomes
$$X(s) = S + W(s).$$
Then, for $j = 1, \ldots, m$,
$$X(s_j) - X(s_0) = W(s_j) - W(s_0) \sim \mathcal{GP},$$
so we can exploit the Gaussianity of the multivariate distribution of
$$(X(s_1) - X(s_0), \ldots, X(s_m) - X(s_0))$$
to estimate the parameters of the process $W(s)$.

- If $S = 0$, model (1) becomes
$$X(s) = RW(s).$$
We can note that, for $j = 1, \ldots, m$,
$$\frac{X(s_j)}{X(s_0)} = \frac{W(s_j)}{W(s_0)}$$
and we can get the analytical form of the multivariate density of the vector
$$\left(\frac{X(s_1)}{X(s_0)}, \ldots, \frac{X(s_m)}{X(s_0)}\right)$$
and use it to estimate the parameters of the process $W(s)$.

- In the general case (1), we can note that, for $j = 2, \ldots, m$,
$$\frac{X(s_j) - X(s_0)}{X(s_1) - X(s_0)} = \frac{W(s_j) - W(s_0)}{W(s_1) - W(s_0)}$$
and we can get the analytical form of the multivariate density of the vector
$$\left(\frac{X(s_2) - X(s_0)}{X(s_1) - X(s_0)}, \ldots, \frac{X(s_m) - X(s_0)}{X(s_1) - X(s_0)}\right)$$
and use it to estimate the parameters of the process $W(s)$.

3.2 Inference for the parameters of S and R

If we denote as $\bar{X}$ and $\bar{W}$ the average over the spatial location of the processes $X(s)$ and $W(s)$, we can note that

$$\bar{W} = \frac{1}{m+1}\sum_{j=0}^{m} W(s_j) \sim N\left(0, \frac{1}{(m+1)^2}\mathbf{1}^\top \Sigma \mathbf{1}\right). \tag{3}$$

Then, for fixed values of the parameters of S and R, we can get an estimate $\hat{\Sigma}$ of the covariance matrix Σ of the Gaussian process $W(s)$ as explained in 3.1 and plug it into equation (3) to get the (univariate) distribution of $\bar{X} = S + R\bar{W}$. We can then find the values of the parameters of S and R that minimize a goodness-of-fit criterion for the empirical distribution function of $\bar{X}$ with respect to the theoretical one computed above, as in Luceño (2006).

4 Data application

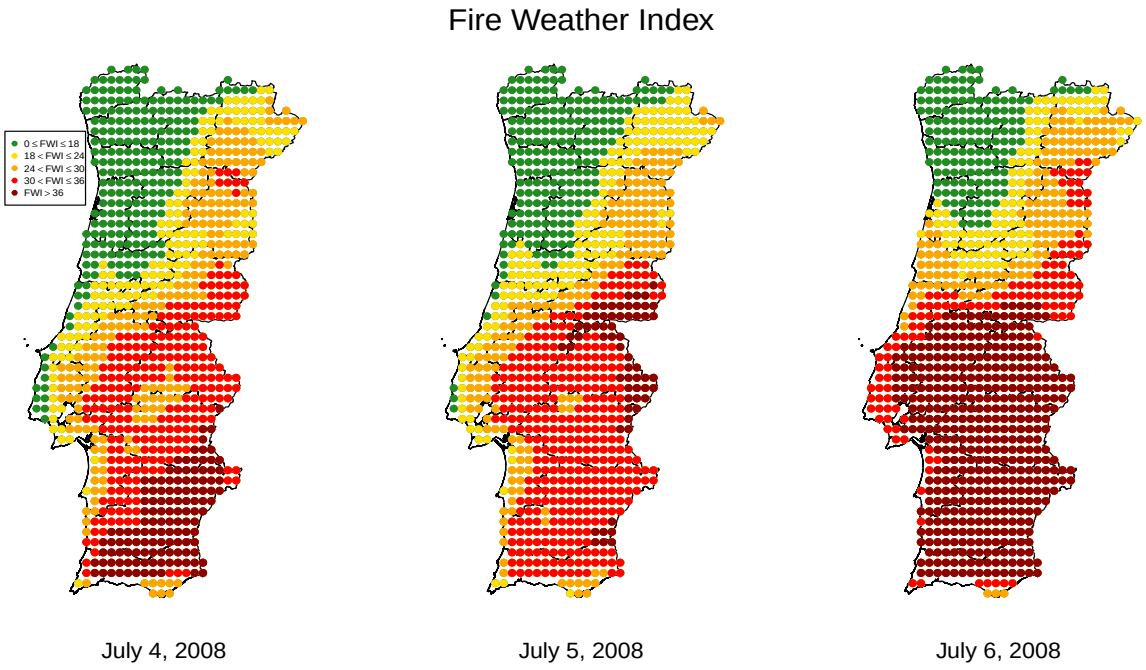


Figure 1: Three consecutive days of observed FWI over Portugal.

ERA5-Land is a large hourly global dataset of various weather variables on a longitude-latitude grid of size $0.1° \times 0.1°$. It is obtained by assimilating observation data (weather stations, radar, satellite...) into a physical climate model. Data are available for the period from 1940 to today, and are routinely used as a realistic approximation of the true weather of

the past. Following common practice, we here refer to them as "observations", even though producing them has required using a model. We apply the random field models we propose to a dataset of daily Fire Weather Index (FWI) computed from ERA5-Land data on the whole domain of continental Portugal for recent years. The dataset includes 20 years of daily observations on 947 spatial locations. For illustration of the data, the maps in Figure 1 show FWI values observed for three consecutive days in summer 2008.

Bibliographie

Engelke, S., Opitz, T., and Wadsworth, J. (2019). Extremal dependence of random scale constructions. *Extremes*, 22(4):623–666.

Huser, R., Opitz, T., and Thibaud, E. (2017). Bridging asymptotic independence and dependence in spatial extremes using gaussian scale mixtures. *Spatial Statistics*, 21:166–186.

Huser, R. and Wadsworth, J. L. (2022). Advances in statistical modeling of spatial extremes. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(1):e1537.

Krupskii, P., Huser, R., and Genton, M. G. (2018). Factor copula models for replicated spatial data. *Journal of the American Statistical Association*, 113(521):467–479.

Luceño, A. (2006). Fitting the generalized pareto distribution to data using maximum goodness-of-fit estimators. *Computational Statistics & Data Analysis*, 51(2):904–917.

Naveau, P., Huser, R., Ribereau, P., and Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4):2753–2769.

Simulation of extreme functionals in meteoceanic data

N. Gorse[1] O.Roustant[1] J.Rohmer[2] D.Idier[2]

[1] *IMT, UMR5219 CNRS, INSA, F-31077 Toulouse cedex 4, France, gorse@insa-toulouse.fr, roustant@insa-toulouse.fr*
[2] *BRGM, 3 avenue Claude Guillemin, F-45060 Orléans cedex 2, France, j.rohmer@brgm.fr, d.idier@brgm.fr*

Short summary We are interested in the influence of meteoceanic conditions on coastal flooding. We specifically work on wave height and focus on extreme observations in the functional domain. Yet, our observations do not meet standard hypotheses as they are dependent and have short-tailed distributions. We propose a three-stage methodology. Firstly, we introduce an autoregressive model in order to remove the temporal dependence between cycles and show that its residuals satisfy the required assumptions. Then, we show how to combine existing techniques related to Pareto processes to simulate extreme residuals. Finally, applying the reverse transformations, we simulate new extreme conditions over tidal cycles. This step depends on an initial time series X_0 that can be chosen to tune the desired level of extremes. The developments are applied to the site of Gâvres in the south of the French Brittany.

Keywords Extreme values, Pareto processes, Regular variations, Autoregressive models

1 Context

The role of meteoceanic conditions is crucial in coastal flooding. For instance, the co-occurence of high wind speed, important surge and wave height may yield large impacts inland. Understanding extreme conditions can bring a new insight into the severity of storms. Our objective is to simulate new observations that have the same behaviour as the observed conditions but are extrapolated towards high values.
The motivation is to enhance the modelling of coastal flooding by a design of experiments (Rohmer et al. 2022). The challenge in our case is that the inputs of such numerical models are functionals (time series) and our objective is to generate extreme observations.
The forcing conditions are defined as the surge, the wind and wave characteristics. We work specifically on the site of Gâvres in French Brittany and use a database based on the paper of (Idier et al. 2020). This small town being located in a macro-tidal area, we focus on meteoceanic conditions occurring (+/-)3h around the high tide (with a fixed time step of 10 minutes). As a consequence, we are interested in the evolution of forcing conditions over tidal cycles and our database consists of time series of length 37 at different events. We will use the notation X_M^t to describe the value obtained at time t for the M^{th} tidal cycle. Thus, $X_M^1, \ldots, X_M^{37}$ represents the time series corresponding to the M^{th} tidal cycle.

2 State-of-the-art

We focus here on the univariate case with an analysis of the significant wave height (Hs). We work in the domain of functional extremes (Dombry and Ribatet 2015; Fondeville and Davison 2022) with independent and identically distributed observations.
We use the "peak-over-threshold" approach (Coles et al. 2001) to define extreme individuals. In this field of research, one must choose a function ℓ which takes a time series as entry and returns a scalar value. It represents the severity of a tidal cycle and is known as a cost functional (Dombry and Ribatet 2015) or a risk functional (Fondeville and Davison 2022). A time series is considered as extreme if the corresponding value of ℓ exceeds some threshold.
It is often assumed that the time series is regularly varying (Hult and Lindskog 2005), implying heavy-tailed marginal distributions. Extreme time series are then associated to the class of Pareto processes (Dombry and Ribatet 2015). Notice that this approach can be generalised to the other types of extremes represented by short-tailed and light-tailed distributions (Fondeville and Davison 2022; Ferreira and Haan 2014).

3 Overview

Our observations do not meet the standard assumptions of independence and regular variation. The first step of our method thus consists in a *"whitening"* pre-processing of the data. More precisely, we extract winter time series and impose a minimal duration between each event. Then, we introduce an autoregressive model accounting for the temporal dependence *between* tidal cycles while preserving the dependence *within* each one of them (and hence the functional nature of the data). We then consider the residuals obtained.
In a second step, we show how to combine the approach of (Opitz, Allard, and Mariethoz 2021) and the work of (Clémençon, Huet, and Sabourin 2024) to simulate new extreme residuals. This involves a marginal transformation to meet the assumption of regular variations.
Finally, in a third step, we apply the reverse transformations and show how to simulate extreme time series. This step depends on an initial time series X_0, that can be chosen to tune the desired level of extremes. For instance, choosing X_0 among the non-extreme observations will produce simulations that mimic the extreme ones, whereas choosing X_0 among the extreme observations will produce more extreme simulations *(uplifting)* corresponding to consecutive extreme tidal cycles.

Step 1. Pre-processing and whitening. We do not have independent series as the first measure of a tidal cycle is for example correlated with its counterpart of the previous observation. A seasonality is also present as extreme values are for instance more frequent in winter.
To ensure independence, we select a subset of winter observations by keeping only 1 tidal cycle out of 3. We then account for temporal dependence between tidal cycles by using, for each $t = 1, \dots, 37$, an autoregressive model of order 1 (where M is the tidal cycle index) :

$$X_M^t = \beta_0 + \beta_1 X_{M-1}^t + \epsilon_M^t, \qquad M = 1, 2, \dots \tag{1}$$

Graphical diagnostics (ACF, PACF) and statistical tests (Pearson) show that (for all t) the residuals ϵ_M^t can be considered as non-correlated. Furthermore, we also analyse the extremal dependence. We use the test of (Thomas 2001) which allows to conclude that the residuals ϵ_M^t are asymptotically independent.

Step 2. Extreme functionals. In this step, we work on the residuals ϵ_M^t. For a given t, we estimate the law of extreme values and reject the heavy-tail hypothesis. Such conclusion has also been observed in the literature for other meteoceanic variables such as wind speed (Palutikof et al. 1999). Thus, in order to get back to the heavy-tail framework, we apply the marginal transformation T described in (Opitz, Allard, and Mariethoz 2021). The transformed residuals $T(\epsilon_M^t)$ then follow a Frechet distribution. Then, we consider for $T(\epsilon_M)$ the framework of (Dombry and Ribatet 2015). The functional data are modeled with a *polar decomposition* of the form

$$f = \ell(f) \times A(f). \tag{2}$$

The radius $\ell(f)$, here chosen as the L^2 norm of f, allows to define extreme functionals with a scalar threshold u. Under the assumption of heavy-tailed distributions, the two components $(\ell(f), A(f))$ become independent when u goes to infinity

$$\ell(f) \perp\!\!\!\perp A(f) \mid \ell(f) > u, \quad u \to +\infty \tag{3}$$

and (f/u) converges in law to a ℓ-Pareto process.

Step 3. Simulation. Simulation of extreme time series can be done from the polar decomposition (2). Simulating the radius $\ell(f)$ is straightforward as its asymptotic distribution is explicit (a Pareto distribution). However, simulating the function $A(f)$ is much more challenging. Here, we follow (Clémençon, Huet, and Sabourin 2024) by simulating the vector of coefficients of a principal component analysis (PCA). The law of that vector is modeled with copulas. As a result, the simulated time series $T(\tilde{\epsilon})$ are approximately drawn from the ℓ-Pareto process of step 2.
Finally, using the reverse transformation T^{-1} and the autoregressive models (1), we obtain simulated time series for $(\tilde{X}^t)_{t=1,\dots,37}$, depending on an initial time series $(X_0^t)_{t=1,\dots,37}$ with

$$\tilde{X}^t = \beta_0 + \beta_1 X_0^t + \tilde{\epsilon}^t \tag{4}$$

4 Numerical results

In this section, we put in action our methodology to provide simulations of extreme time series, using (4). This procedure depends on an initial time series X_0. We first compare two choices of X_0: one corresponding to the extreme storm event Johanna (08/03/2008 − 10/03/2008) and another one chosen among non-extreme events in the sense that $\ell(T(\epsilon_M)) < u$ (see step 2). The result is presented in Figure 1. We see that the choice of X_0 influences the values and the shape of $\tilde{X}$. We also see on the left panel that we obtain very large values which may be unrealistic. These levels represent the effect of having consecutive extremes.

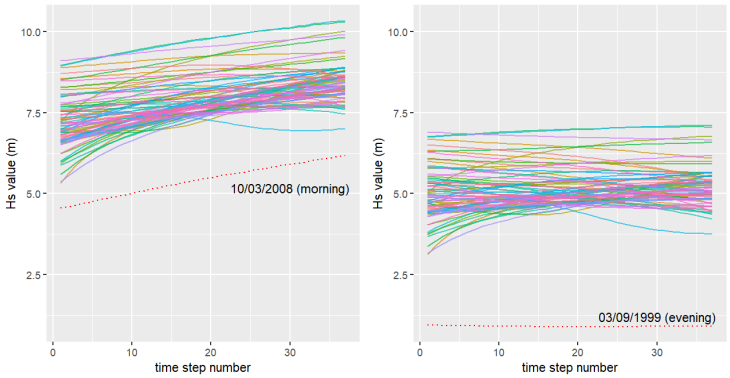


Figure 1: Simulated time series obtained with two different types of X_0 (dotted line) selected among extreme observations (left panel) or non-extreme ones (right panel).

Our aim is to obtain realistic extreme simulations. Thus, from now on we draw randomly each X_0 from the subset of non-extreme observations in the sense that $\ell(X_0) < 3.4$m and, for each realization, we use (4) to produce extreme simulations $\tilde{X}$. We assess the quality of our method by checking that simulations and extreme observations share common behaviours. We use several tools.

First of all, they should have shape similarities. As the simulations extrapolate the L^2 norm of the observations, we use a PCA on the normalized data $X/\ell(X)$. For both simulations and observations, we further standardize by removing the mean and dividing by the standard deviation of the observations. For illustration, we select two samples of size 78 (20% of the database of extreme observations) (Figure 2).

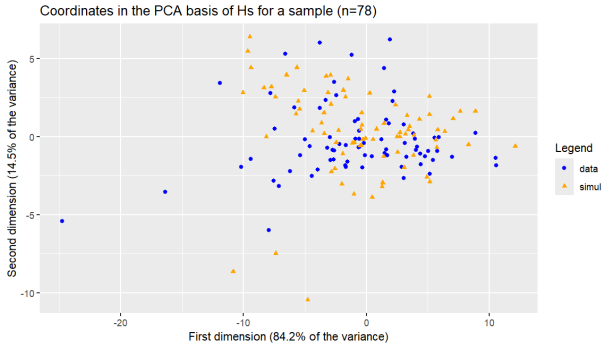


Figure 2: PCA coordinates for Hs

The simulated time series (orange triangles) generally behave the same way as the extreme observations (blue dots) as their respective coordinates are close. Yet, it is clearer for the first dimension, which explains the largest proportion of variability (here almost 85%). The Kolmogorov-Smirnov test confirms these remarks as simulations' and observations' coordinates follow the same law in the first dimension but do not in the second dimension.

Besides, we compare the temporal dependence within each tidal cycle between observations and simulations with a focus on extremal dependence. We use the extremogram (Davis

and Mikosch 2009, Fondeville and Davison 2022) to measure this dependence. The function obtained in the simulations (not shown here) is close to the data estimator.

Moreover, the density obtained at each time t should be an extrapolation of what we observe in the database. We compare the quantiles of the simulations, using the law of extremes for X_M^t, with the empirical quantiles of the observations. Following (Solari et al. 2017), the results are presented in term of return level and return period (Figure 3). The link between the quantile level p and the return period T is given by the formula

$$T = \frac{1}{\text{npy}(1-p)}$$

where npy is the average number of observations per year. Thus, looking at Figure 3 (left panel), if we observe one day a significant wave height of 7.5m at tidal peak ($t = 19$ for every cycle), we will wait in average 20 years before seeing an event reaching this extreme level. The largest simulated levels are located within the bounds of the confidence interval. It induces that our simulations look similar to the extreme observations. They can be considered as an extrapolation of the observations towards high values. Yet, we see that some simulated levels are slightly above the confidence band. The choice of X_0 may intervene here as we can have individuals with too large values.

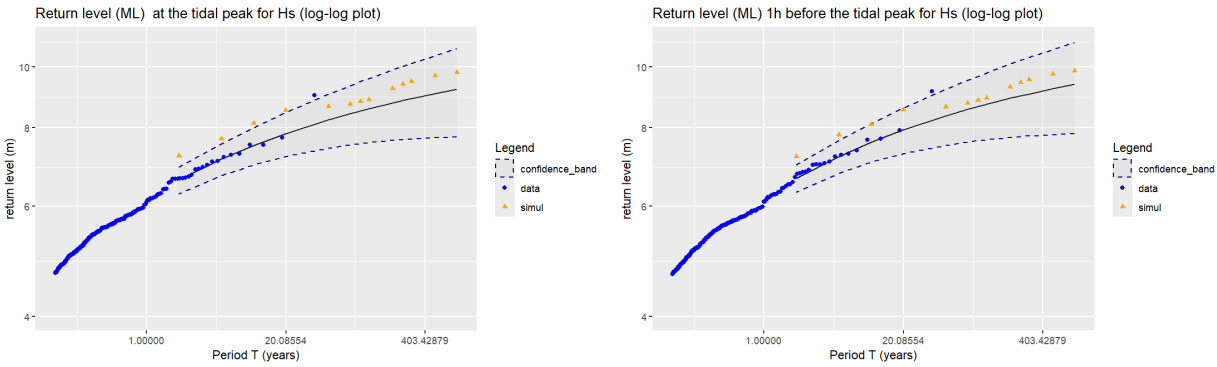


Figure 3: Return level for simulations (orange triangles) and observations (blue dots), for two different values of t. Left panel: $t = 19$ (tidal peak), right panel: $t = 13$ (one hour before). The dotted lines represent 99% confidence bands.

To conclude, we evaluate the consistency of our simulations with the extreme observations by using a machine learning model. We use a classification problem where the target variable equals 1 if the input is a simulated time series and 0 otherwise. If simulations behave as extreme observations, such model should hardly identify the simulated time series, and the rate of correct prediction should be around 50%. Applying this procedure on ϵ_M, we obtain a rate of 44% with a logistic model and 54% with a SVM if we use the radial kernel, showing that the simulated $\tilde{\epsilon}$ look similar to the residuals associated to observations. Doing the same on X_t, we obtain respectively a rate of 66% and 64%, showing a little difference. This may be improved for another choice of X_0.

Bibliography

Clémençon, S., Huet, N., and Sabourin, A. (2024). "Regular variation in Hilbert spaces and principal component analysis for functional extremes". In: *Stochastic Processes and their Applications* 174, p. 104375.

Coles, S. et al. (2001). *An introduction to statistical modeling of extreme values*. Vol. 208. Springer.

Davis, R. A. and Mikosch, T. (2009). "The extremogram: A correlogram for extreme events". In: *Bernoulli* 15.4, pp. 977–1009.

Dombry, C. and Ribatet, M. (2015). "Functional regular variations, Pareto processes and peaks over threshold". In: *Statistics and its Interface* 8.1, pp. 9–17.

Ferreira, A. and Haan, L. de (2014). "The generalized Pareto process; with a view towards application and simulation". In: *Bernoulli*, pp. 1717–1737.

Fondeville, R. de and Davison, A. C. (2022). "Functional peaks-over-threshold analysis". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.4, pp. 1392–1422.

Hult, H. and Lindskog, F. (2005). "Extremal behavior of regularly varying stochastic processes". In: *Stochastic Processes and their applications* 115.2, pp. 249–274.

Idier, D. et al. (Mar. 2020). "Coastal flood: a composite method for past events characterisation providing insights in past, present and future hazards—joining historical, statistical and modelling approaches". In: *Natural Hazards* 101. DOI: 10.1007/s11069-020-03882-4.

Opitz, T., Allard, D., and Mariethoz, G. (2021). "Semi-parametric resampling with extremes". In: *Spatial Statistics* 42, p. 100445.

Palutikof, J. P. et al. (1999). "A review of methods to calculate extreme wind speeds". In: *Meteorological applications* 6.2, pp. 119–132.

Rohmer, J. et al. (2022). "Partitioning the contributions of dependent offshore forcing conditions in the probabilistic assessment of future coastal flooding". In: *Natural Hazards and Earth System Sciences* 22.10, pp. 3167–3182.

Solari, S. et al. (2017). "Peaks O ver T hreshold (POT): A methodology for automatic threshold estimation using goodness of fit p-value". In: *Water Resources Research* 53.4, pp. 2833–2849.

Thomas, M. (2001). *Statistical analysis of extreme values: with applications to insurance, finance, hydrology, and other fields*. Birkhäuser Verlag.

Spatio-temporal modeling of urban extreme rainfall events at high resolution

Chloé Serre-Combe [1] & Nicolas Meyer [2] & Thomas Opitz [3] & Gwladys Toulemonde [4]

[1] *Univ. Montpellier, CNRS, IMAG, Inria Lemon, France,* *chloe.serre-combe@umontpellier.fr*
[2] *Univ. Montpellier, CNRS, IMAG, Inria Lemon, France,* *nicolas.meyer@umontpellier.fr*
[3] *INRAE, BioSP, Avignon, France,* *thomas.opitz@inrae.fr*
[4] *Univ. Montpellier, CNRS, IMAG, Inria Lemon, France,* *gwladys.toulemonde@umontpellier.fr*

Résumé. La modélisation des précipitations est essentielle pour analyser les risques d'inondation. Nous proposons un cadre intégrant un réseau de pluviomètres de Montpellier et des données radar de Météo-France pour modéliser la distribution et la dépendance spatio-temporelle des précipitations à haute résolution temporelle et fine échelle spatiale. La loi de Pareto généralisée étendue (EGPD) est utilisée pour représenter en univarié simultanément les précipitations modérées et extrêmes sans choix de seuil, simplifiant ainsi l'estimation des paramètres. La dépendance spatio-temporelle est capturée par un processus de r-Pareto avec une structure de dépendance gaussienne et un variogramme spatio-temporel non séparable avec advection, permettant des simulations plus réalistes de précipitation. Une approche de vraisemblance composite, basée sur des indicateurs de dépassement conjoints bivariés, est utilisée pour estimer les paramètres du variogramme. Validée par simulations, cette approche sera appliqué aux données de Montpellier pour ensuite servir de base à la construction d'un générateur stochastique de précipitations.

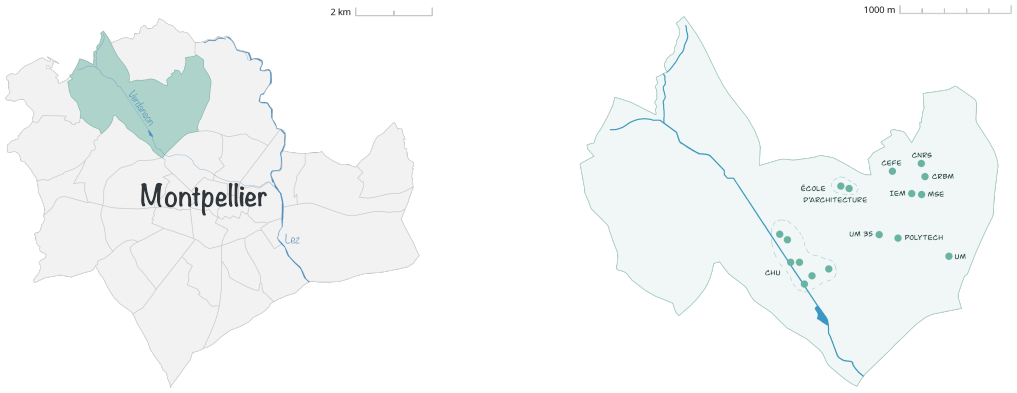
Mots-clés. Théorie des valeurs extrêmes, EGPD, haute résolution spatio-temporelle, processus de r-Pareto, advection

Abstract. Precipitation modeling is of great interest for flood risk analysis. We propose a framework integrating a rain gauge network in Montpellier and Météo-France radar data to model the distribution and spatio-temporal dependence of rainfall at high temporal resolution and fine spatial scale. The Extended Generalized Pareto Distribution (EGPD) is used for univariate modeling, capturing both moderate and extreme rainfall without threshold selection, simplifying parameter estimation. Spatio-temporal dependence is modeled using an r-Pareto process with a Gaussian dependence structure and a non-separable variogram with advection, enabling more realistic precipitation simulations. A composite likelihood approach based on bivariate joint exceedance indicators is used for variogram parameter estimation. Validated through simulations, this approach will be applied to Montpellier data as a basis for developing a stochastic precipitation generator.

Keywords. Extreme value theory, EGPD, high spatio-temporal resolution, r-Pareto process, advection

1 Context and study area

Montpellier, with its Mediterranean climate and urbanization, faces heavy rainfall events leading to flooding. Its proximity to the Cévennes mountains exacerbates flood risks, highlighting the need for precise rainfall simulations for flood assessment and urban planning. The localized nature of these events underscores the importance of high-resolution spatio-temporal rainfall modeling.



(a) City district division (b) Rain gauges

Figure 1: Location of the 17 rain gauges of the OMSEV rain gauges network

The study focuses on the OMSEV[1] rain gauge network, which includes 17 stations in the Verdanson catchment, a tributary of the Lez (see Figure 1). Rainfall data, provided by HydroScience Montpellier (Finaud-Guyot et al., 2023), spans from 2019 to 2022 with minute-level measurements, aggregated into 5-minute intervals to reduce errors while maintaining high temporal resolution. Due to this fine scale, only 1.2% of the data consists of non-zero values. The station spacing ranges from 77 to 1531 meters, highlighting the fine spatial resolution. Additionally, the COMEPHORE dataset from Météo France (Tabary et al., 2012) will be considered. This is reanalysis data from 1997 to 2021, providing hourly cumulative rainfall at a 1 km^2 resolution by combining radar and gauge measurements for better spatial consistency. The main objective is to model rainfall's spatio-temporal variability in the Montpellier area and simulate realistic rainfall events using a stochastic precipitation generator, integrating both datasets.

2 Modeling univariate margins

To model the marginal distribution of positive rainfall, the Extended Generalized Pareto Distribution (EGPD) is used. This distribution avoids explicit threshold selection and efficiently captures both moderate and extreme events. The EGPD is a transformation of the Generalized Pareto Distribution (GPD), which models threshold exceedances with shape parameter

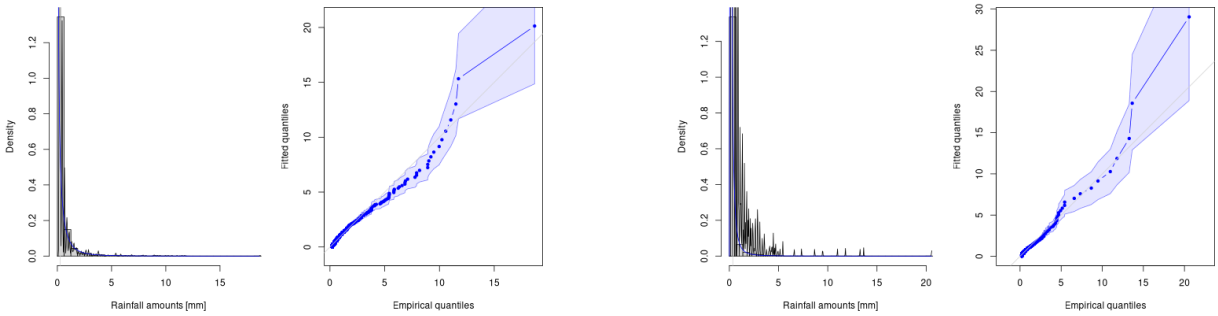
[1]Observatoire Montpelliérain et au Sud de l'Eau dans la Ville

ξ and a scale parameter σ_u, with u a fixed threshold u. The cumulative distribution function is approximated, for a rainfall intensity $X_{\boldsymbol{s}}$ at location $\boldsymbol{s} \in \mathbb{R}^2$, as

$$\mathbb{P}\left(X_{\boldsymbol{s}} - u > y | X_{\boldsymbol{s}} > u\right) \approx \overline{H}_{\xi}\left(\frac{y}{\sigma_u}\right) = \begin{cases} \left(1 + \xi \frac{y}{\sigma_u}\right)_{+}^{-1/\xi} & \text{if } \xi \neq 0\,, \\ e^{-\frac{y}{\sigma_u}} & \text{if } \xi = 0\,, \end{cases}$$

with $a_+ = \max(a, 0)$. A continuous function $G(x) = x^{\kappa}$ is used to modify the GP distribution, where κ controls the moderate part of the distribution. Then, the upper tail still follows a GP distribution, while the lower tail is based on a GP distribution for negative values with an upper bound of 0, leading to a power-law behavior.

The EGPD is fitted to the OMSEV data, with small left-censoring adjustments applied at each station to improve model performance, according the Root Mean Square Error (RMSE) as a goodness-of-fit measure (Haruna et al., 2023). Good fits are obtained for all rain gauges, as shown in the quantile-quantile plot in Figure 2. So, the EGPD model is a good candidate for modeling the marginal distribution of rainfall at each station.



(a) CRBM (b) CNRS

Figure 2: EGPD fitting on two rain gauges of the OMSEV network with the density estimation (left) and the quantile-quantile plot (right).

3 Spatio-temporal dependence modeling

Let $\boldsymbol{X} = \{X_{\boldsymbol{s},t} \mid (\boldsymbol{s}, t) \in \mathcal{S} \times \mathcal{T}\}$ be a strictly stationary isotropic random process that represents the rainfall field over the study area. The spatio-temporal dependence is modeled by an r-Pareto process, as described by de Fondeville and Davison, 2018, which is an alternative to max-stable processes for threshold exceedances.

Let's consider a nonnegative and 1-homogeneous risk function $r(\boldsymbol{X}) = X_{\boldsymbol{s}_0,t_0}$. As described in Dombry et al., 2024, the r-exceedances of $\boldsymbol{X}$, defined by $\{X_{\boldsymbol{s},t} \mid r(\boldsymbol{X}) > u\}$, converge in distribution to an r-Pareto process $\boldsymbol{Y} = \{Y_{\boldsymbol{s},t}, (\boldsymbol{s}, t) \in \mathcal{S} \times \mathcal{T}\}$ when u goes to infinity. This process is defined by a Gaussian process $\boldsymbol{W}$, associated with a variogram γ, that captures the spatial and temporal dependence. The r-Pareto process is given by

$$Y_{\boldsymbol{s},t} = uR_{\boldsymbol{s},t} e^{W_{\boldsymbol{s},t} - W_{\boldsymbol{s}_0,t_0} - \gamma_r(\boldsymbol{s},t)}\,,$$

where $R_{\boldsymbol{s},t}$ follows a univariate standard Pareto distribution, and $\gamma_r(\boldsymbol{s},t) = \gamma(\boldsymbol{s}-\boldsymbol{s}_0, t-t_0)$. Our goal is to model the dependence structure of $\boldsymbol{Y}$, which is directly determined by the dependence structure of $\boldsymbol{W}$, given by the spatio-temporal variogram γ. This is achieved by examining the space-time extremogram, which measures the extreme spatio-temporal dependence by analyzing the simultaneous excesses for pairs of locations with spatial and temporal lags. Let $\Lambda_{\mathcal{S}} \subset \mathbb{R}^2$ and $\Lambda_{\mathcal{T}} \subset \mathbb{R}_+$ be sets of spatial and temporal lags. The extremal dependence measure, or extremogram, is defined for $\boldsymbol{h} \in \Lambda_{\mathcal{S}}, \tau \in \Lambda_{\mathcal{T}}$ as (Coles et al., 1999):

$$\chi(\boldsymbol{h},\tau) = \lim_{q\to 1} \chi_q(\boldsymbol{h},\tau)\ , \quad \text{with} \quad \chi_q(\boldsymbol{h},\tau) = \mathbb{P}(X^*_{\boldsymbol{s},t} > q \mid X^*_{\boldsymbol{s}+\boldsymbol{h},t+\tau} > q)\,,$$

where $q \in [0,1[$ and for any $\boldsymbol{s} \in \mathcal{S}$ and any $t \in \mathcal{T}$, $X^*_{\boldsymbol{s},t}$ is the uniform rank transformation of $X_{\boldsymbol{s},t}$. For r-Pareto processes, the r-extremogram χ_r can be defined as the extremogram of the r-exceedances of our rainfall process $\boldsymbol{X}$, which will be directly the marginals of the process. The dependence structure of r-Pareto processes, which is inherited from the Gaussian framework, leads to a relationship between the r-variogram and the r-extremogram (Davis et al., 2013 and Coles et al., 1999):

$$\chi_r(\boldsymbol{h},\tau) = 2\left(1-\phi\left(\sqrt{\frac{1}{2}\gamma_r(\boldsymbol{h},\tau)}\right)\right),\ \boldsymbol{h} \in \Lambda_{\mathcal{S}},\ \tau \in \Lambda_{\mathcal{T}}.$$

where ϕ is the standard normal distribution function.

4 Estimation of the variogram parameters

4.1 Variogram structure

To model spatial-temporal dependencies in rainfall, we use a non-separable variogram of a fractional Brownian motion, which accounts for advection, which is the horizontal transport of air masses. The variogram, incorporating a velocity vector $\boldsymbol{V} \in \mathbb{R}^2$, is given by:

$$\gamma(\boldsymbol{h},\tau) = 2\left(\beta_1\|\boldsymbol{h}-\tau\boldsymbol{V}\|^{\alpha_1} + \beta_2|\tau|^{\alpha_2}\right), \quad \boldsymbol{h} \in \Lambda_{\mathcal{S}}, \tau \in \Lambda_{\mathcal{T}},$$

with $\beta_1 > 0, \beta_2 > 0, 0 < \alpha_1, \alpha_2 \leq 2$. If $\boldsymbol{V} = \boldsymbol{0}$, the variogram is separable in space and time.

4.2 Composite likelihood approach

The variogram parameters $\boldsymbol{\Theta} = (\beta_1, \beta_2, \alpha_1, \alpha_2, \boldsymbol{V})$ are estimated by maximizing the composite likelihood function. In spatio-temporal processes, the dependence structure of extremes can be characterized by the total number of joint exceedances across site pairs separated by a spatial lag $\boldsymbol{h}$ and a temporal lag τ:

$$K_{\boldsymbol{h},\tau}(u) \ = \sum_{(\boldsymbol{s}_i,\boldsymbol{s}_j)\in\mathcal{N}(\boldsymbol{h})} \sum_{(t_i,t_j)\in\mathcal{N}(\tau)} \mathbb{1}_{\{X_{\boldsymbol{s}_i,t_i}>u, X_{\boldsymbol{s}_j,t_j}>u\}}.$$

where $\mathcal{N}(\boldsymbol{h})$ and $\mathcal{N}(\tau)$ are the sets of all pairs of sites separated by the spatial lag $\boldsymbol{h} \in \Lambda_{\mathcal{S}}$ and all pairs of time indices separated by the temporal lag $\tau \in \Lambda_{\mathcal{T}}$, respectively.

For considered r-Pareto processes, the indicator of joint exceedances becomes $k_{\boldsymbol{s},t} = \mathbb{1}_{\{X_{\boldsymbol{s}_0,t_0}>u, X_{\boldsymbol{s},t}>u\}}$ for $(\boldsymbol{s}, \boldsymbol{s}_0) \in \mathcal{N}(\boldsymbol{h})$ and $(t, t_0) \in \mathcal{N}(\tau)$. Then, since t_0 is fixed, the total number of temporal pairs within temporal lag τ is 1, while the number of spatial pairs corresponds to the number of pairs of sites separated by $\boldsymbol{h}$ from the site $\boldsymbol{s}_0$. When u goes to infinity, the total of joint exceedances follow a binomial distribution, with the number of trials equal to the number of spatial pairs and the probability of success is directly given by the r-extremogram and the marginal probability of exceedance above u, which is $\mathbb{P}(X_{\boldsymbol{s}_0,t_0} > u) = 1$. Then, the composite log-likelihood to maximize is given by

$$l_C(\boldsymbol{\Theta}) \propto \sum_{(\boldsymbol{s}_0,t_0)\in\mathcal{R}} \sum_{(\boldsymbol{s},\boldsymbol{s}_0)\in\mathcal{N}(\boldsymbol{h})} \sum_{(t,t_0)\in\mathcal{N}(\tau)} k_{\boldsymbol{s},t} \log \chi_{r,\boldsymbol{\Theta}}(\boldsymbol{h},\tau) + (1 - k_{\boldsymbol{s},t}) \log(1 - \chi_{r,\boldsymbol{\Theta}}(\boldsymbol{h},\tau)).$$

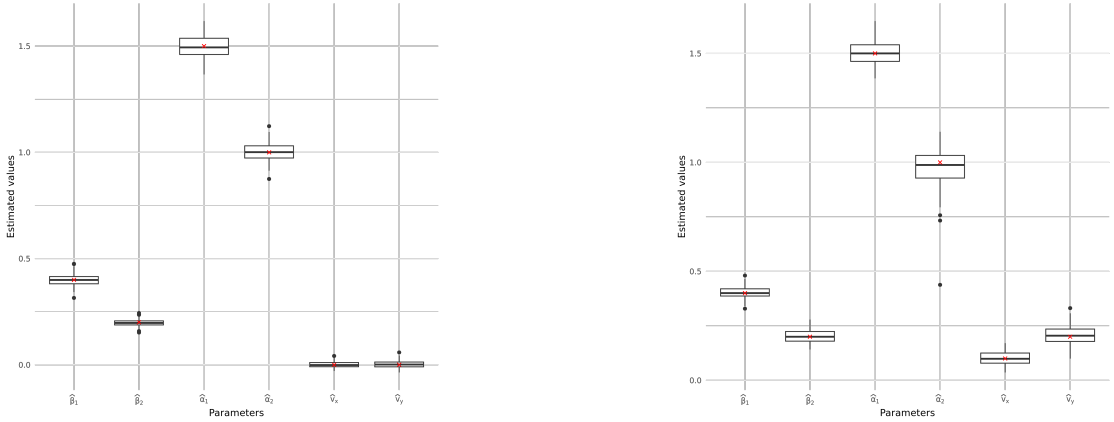
where $\mathcal{R}$ denotes the set of conditioning spatio-temporal points for which there is an exceedance, where no two points $(\boldsymbol{s}, t)$ and (s', t') in $\mathcal{R}$ are within the same spatial and temporal neighbourhood. For real data applications, the optimization is initialized following the approach of Buhl et al., 2019, using a separable variogram and Weighted Least Squares estimation of the parameters.

4.3 Model validation on simulations

The simulations of r-Pareto processes were performed by adapting the methodology of the exact simulation of spatial max-stable processes proposed by Dombry et al., 2016 and the algorithm of Leber, 2015 for simulating spatio-temporal Brown-Resnick processes. The algorithm was modified to incorporate a Pareto random variable and it simulates only one Gaussian process. Simulations were conducted on a regular grid with a threshold $u = 1$, defining the r-exceedances. The risk function $r(\boldsymbol{X}) = X_{\boldsymbol{s}_0,t_0}$ was used, conditionning the process on the first site and time step. The variogram parameter estimation was conducted on 100 simulated r-Pareto processes with 1000 replicates each, producing excellent results, as illustrated in the boxplots of Figure 3, with constant advection and without advection.

5 Outlook

The proposed approach will be applied to real COMEPHORE data to estimate an advection vector for each extreme episode, where observed wind speed data will be used as a covariate to include event-specific information on advection. The resulting advection will then be used as the constant advection parameter to optimize the other variogram parameters on OMSEV data. Finally, this process will allow the simulation of realistic rainfall events, improving the analysis of flood risks in the Montpellier area. To facilitate its use and reproducibility, we aim to develop it as an `R` package. The code is currently available on GitHub: chloesrcb/generain.



(a) Without advection.

(b) With advection.

Figure 3: Estimations of variogram parameters with maximum likelihood optimization of 100 simulated r-Pareto processes with 1000 replicates each, with 25 sites and 30 time observations. The true parameters are given by red crosses: $\boldsymbol{\Theta} = (0.4, 0.2, 1.5, 1, (0, 0))$ (left) and $\boldsymbol{\Theta} = (0.4, 0.2, 1.5, 1, (0.1, 0.2))$ (right).

References

Buhl, S., Davis, R. A., Klüppelberg, C., & Steinkohl, C. (2019). Semiparametric estimation for isotropic max-stable space-time processes. *Bernoulli*, *25*.

Coles, S., Heffernan, J., & Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, *2*.

Davis, R. A., Klüppelberg, C., & Steinkohl, C. (2013). Max-stable processes for modeling extremes observed in space and time. *Journal of the Korean Statistical Society*, *42*(3).

de Fondeville, R., & Davison, A. C. (2018). High-dimensional peaks-over-threshold inference. *Biometrika*, *105*(3).

Dombry, C., Legrand, J., & Opitz, T. (2024). *Pareto processes for threshold exceedances in spatial extremes*.

Dombry, C., Engelke, S., & Oesting, M. (2016). Exact simulation of max-stable processes. *Biometrika*, *103*(2), 303–317.

Finaud-Guyot, P., Guinot, V., Marchand, P., Neppel, L., Salles, C., & Toulemonde, G. (2023). Rainfall data collected by the HSM urban observatory (OMSEV).

Haruna, A., Blanchet, J., & Favre, A.-C. (2023). Modeling intensity-duration-frequency curves for the whole range of non-zero precipitation: A comparison of models. *Water Resources Research*, *59*(6).

Leber, D. (2015). *Comparison of simulation methods of brown-resnick processes*.

Tabary, P., Dupuy, P., L'Henaff, G., Gueguen, C., Moulin, L., & Laurantin, O. (2012). A 10-year (1997—2006) reanalysis of quantitative precipitation estimation over france: Methodology and first results. *IAHS-AISH publication*, (351).

Extrêmes et risques

Estimation of extreme risk measures with neural networks

Michaël Allouche [1], Stéphane Girard [2] & Emmanuel Gobet [3]

[1] *Kaiko - Quantitative Data, 2 rue de Choiseul, 75002 Paris, France;*
[2] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France; stephane.girard@inria.fr.*
[3] *Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique and CNRS, Institut Polytechnique de Paris, Route de Saclay, 91128 Palaiseau Cedex, France; emmanuel.gobet@polytechnique.edu.*

Résumé. Nous proposons de nouvelles paramétrisations de réseaux de neurones afin d'estimer des mesures de risque extrêmes tels que les moments conditionnels de queue, dans des contextes de lois à queues lourdes. L'estimateur proposé est capable d'extrapoler dans les queues de distribution grâce à une extension de la condition de second-ordre à un ordre arbitraire. Le taux de convergence de l'erreur uniforme entre le log-moment conditionnel de queue et son approximation par le réseau de neurones est établi. La performance en pratique de l'estimateur obtenu est comparée sur des données simulées à celle de concurrents issus de la théorie des valeurs extrêmes et incluant une réduction du biais. Il est démontré que notre méthode les surpasse dans les situations difficiles où les autres estimateurs échouent presque tous. Enfin, l'estimateur par réseaux de neurones est testé sur des données réelles afin d'étudier le comportement de rendements extrêmes de crypto-monnaies.

Mots-clés. Statistique des valeurs extrêmes, mesure de risque, lois à queue lourde, réseau de neurones.

Abstract. We propose new parameterizations for neural networks in order to estimate extreme risk measures, such as conditional tail moments, in heavy-tailed settings. The proposed neural network estimator is able to extrapolate in the distribution tails thanks to an extension of the usual extreme-value second-order condition to an arbitrary order. The convergence rate of the uniform error between the log-conditional tail moment and its neural network approximation is established. The finite sample performance of the neural network estimator is compared to bias-reduced extreme-value competitors on simulated data. It is shown that our method outperforms them in difficult heavy-tailed situations where other estimators almost all fail. Finally, the neural network estimator is tested on real data to investigate the behavior of cryptocurrency extreme loss returns.

Keywords. Extreme value theory, risk measure, heavy-tailed distribution, neural networks.

1 Introduction

The Value at Risk (VaR) is the quantile at a given confidence level $(1-\alpha)$:

$$\texttt{VaR}(1-\alpha) = q(1-\alpha) = \inf\{x \in \mathbb{R},\, F(x) \geq 1-\alpha\},$$

where $\alpha \in (0,1)$ and with F the cumulative distribution function of the random variable X of interest (X is usually the opposite of a loss and α is close to 0). Besides, the ES is defined as the average of the quantile function above the confidence level $(1-\alpha)$:

$$\texttt{ES}(1-\alpha) = \frac{1}{\alpha}\int_0^\alpha q(1-u)\,\mathrm{d}u.$$

When the distribution of X is continuou, $\texttt{ES}(1-\alpha)$coincides with the conditional expectation of X given that it exceeds the VaR $q(1-\alpha)$:

$$\texttt{ES}(1-\alpha) = \mathbb{E}\left(X \mid X > q(1-\alpha)\right),$$

and it is referred to as Conditional Tail Expectation (CTE). In this work, we consider more generally Conditional Tail Moments (CTMs) of order $p > 0$, defined in [5] as

$$\texttt{CTM}_p(1-\alpha) = \mathbb{E}\left(X^p \mid X > q(1-\alpha)\right) = \frac{1}{\alpha}\int_0^\alpha q^p(1-u)\,\mathrm{d}u \tag{1}$$

under the assumption of a continuous distribution. Clearly, $\texttt{CTM}_1 = \texttt{ES}$. Nonparametric estimators have been introduced in [3, 4] based on the empirical counterpart of (1):

$$\widehat{\texttt{CTM}}_p(1-\alpha) = \frac{1}{\lfloor n\alpha \rfloor}\sum_{j=1}^{\lfloor n\alpha \rfloor} X_{n-j+1,n}^p \text{ with } \widehat{\texttt{ES}}(1-\alpha) = \widehat{\texttt{CTM}}_1(1-\alpha), \tag{2}$$

and where $X_{1,n} \leq \cdots \leq X_{n,n}$ denote the order statistics associated with the initial n-sample $\{X_1, \ldots, X_n\}$. We consider extreme risks, *i.e.* the situation where α is so small that the VaR is larger than the maximal observation: $q(1-\alpha) > X_{n,n}$. This is asymptotically the case when $n\alpha$ is small, and in this challenging situation, the naive estimator (2) cannot be used to estimate the extreme ES or CTM. To overcome this problem, we propose an extrapolation formula dedicated to heavy-tailed distributions and leveraging extreme-value theory.

2 An approximation result

Let $\{X_1, \ldots, X_n\}$ be an independent sample from an unknown cumulative distribution function F assumed to be continuous. The associated order statistics are denoted by $X_{1,n} \leq \cdots \leq X_{n,n}$. We address the estimation of $\texttt{ES}(1-\alpha_n)$ at an extreme level α_n *i.e.* such that $n\alpha_n \to 0$ as $n \to \infty$. We moreover focus on heavy-tailed distributions whose survival function $(1-F)$ is regularly-varying with index $-1/\gamma$, where $\gamma > 0$ is called the tail-index. Our main assumption can be written as

(**A**) $(1-F) \in \mathcal{RV}_{-1/\gamma}$ with $\gamma > 0$ *i.e.*

$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-1/\gamma}, \text{ for all } x > 0.$$

The index γ tunes the tail heaviness of F: the larger the index, the heavier the right tail. Examples include the (generalized) Pareto, Burr, Fisher, Inverse gamma and Student distributions. Equivalently, the tail quantile function $t > 1 \mapsto U(t) = q(1-1/t)$ can be rewritten as $U(t) = t^{\gamma}L(t)$ where $L \in \mathcal{RV}_0$ is called a slowly-varying function. It is easily shown that, if (**A**) holds with $p\gamma \in (0,1)$, then $\texttt{CTM}_p(1-\alpha)$ is finite for any $\alpha \in (0,1)$. We also consider a J-th order condition ($J \geq 2$), introduced in [10], on the slowly-varying function to control the rate of convergence in (**A**):

(**B$_\mathbf{J}$**) There exist for all $j \in \{2,\ldots,J\}$, functions A_j with $A_j(t) \to 0$ as $t \to \infty$ and an asymptotically constant sign, $|A_j| \in \mathcal{RV}_{\rho_j}$ and $\rho_j \leq 0$ such that

$$\log L(tz) - \log L(t) = \sum_{j=2}^{J} \prod_{\ell=2}^{j} A_\ell(t) R_j(z) + o\left(\prod_{j=2}^{J} A_j(t)\right),$$

as $t \to \infty$ for all $z > 0$, where:

$$R_j(z) := \int_1^z z_2^{\rho_2-1} \int_1^{z_2} z_3^{\rho_3-1} \cdots \int_1^{z_{j-1}} z_j^{\rho_j-1} \,\mathrm{d}z_j \ldots \mathrm{d}z_3 \,\mathrm{d}z_2.$$

This condition refines the initial heavy-tail assumption and is a classical device in extreme-value analysis for extrapolating beyond the observation range. For instance, when $J = 2$, we recover the classical second-order condition [8, Equation (13)]. Moreover, $J = 3$ and $J = 4$ yield back respectively the third-order [6] and fourth-order conditions [7]. For most of heavy-tailed distributions, $A_1, \ldots, A_J$ are power functions, which will be our assumption in the sequel:

(**C$_\mathbf{J}$**) $A_j(t) = c_j t^{\rho_j}$ where $c_j \neq 0$ and $\rho_j < 0$ for $j \in \{2,\ldots,J\}$.

When $J = 2$, (**C$_\mathbf{J}$**) amounts to supposing that the underlying distribution belongs to the Hall-Welsh class [9]. Finally, we shall also need the following assumption.

(**D**) L is differentiable and $|(\log L)'| \in \mathcal{RV}_{\rho_2-1}$ with $\rho_2 < 0$.

Assume (**A**) holds with $p\gamma \in (0,1)$, and consider the slowly-varying function $t \mapsto L(t) = t^{-\gamma}U(t)$. The following two functions are then introduced:

$$(y_1,y_2) \in [0,\infty)^2 \mapsto \varphi(y_1,y_2) = \log L(\exp(y_1+y_2)) - \log L(\exp(y_2)),$$
$$u \in (0,1] \mapsto \Psi_p(u) = \log\left(\int_0^1 w^{-p\gamma} \frac{L^p(1/(wu))}{L^p(1/u)} \,\mathrm{d}w\right).$$

The proposed approach consists in estimating these functions φ and Ψ_p using neural networks (NNs). In [1], we have established in a similar statistical framework that the log-spacings between two quantiles can be written with φ as

$$\log q(1-\alpha) - \log q(1-\delta) = \gamma \log(\delta/\alpha) + \varphi\big(\log(\delta/\alpha), \log(1/\delta)\big).$$

The log-spacings associated with the CTM can be rewritten thanks to φ and Ψ_p.

Proposition 1 *If* $(\mathbf{A})$ *holds with* $p\gamma \in (0,1)$*, then, for all* $(\alpha,\delta) \in (0,1)^2$*:*

$$\begin{aligned}\log \mathtt{CTM}_p(1-\alpha) - \log \mathtt{CTM}_p(1-\delta) &= p\gamma \log(\delta/\alpha) + p\varphi\big(\log(\delta/\alpha), \log(1/\delta)\big) \\ &\quad + \Psi_p(\alpha) - \Psi_p(\delta) \\ &=: g_p(\alpha,\delta).\end{aligned}$$

Proposition 1 allows us to derive an extrapolation principle for the CTM. Indeed, the CTM at an extreme level $1-\alpha_n$ (such that $n\alpha_n \to 0$ as $n \to \infty$) can be deduced from the CTM at an intermediate level $1-\delta_n$ (such that $n\delta_n \to \infty$ as $n \to \infty$) since

$$\mathtt{CTM}_p(1-\alpha_n) = \mathtt{CTM}_p(1-\delta_n)(\delta_n/\alpha_n)^{p\gamma} \exp\big(p\varphi\big(\log(\delta_n/\alpha_n), \log(1/\delta_n)\big) + \Psi_p(\alpha_n) - \Psi_p(\delta_n)\big),$$

or equivalently

$$\mathtt{CTM}_p(1-\alpha_n) = \mathtt{CTM}_p(1-\delta_n)\exp(g_p(\alpha_n,\delta_n)).$$

This kind of extrapolation method is widely used in extreme-value theory, see for instance Weissman estimator [11] dedicated to extreme quantiles. The challenge is to design an extrapolation function g_p in a way that the approximation of $\mathtt{CTM}_p(1-\alpha_n)$ is as stable as possible with respect to the anchor level δ_n. The implementation of this idea requires the estimation of two real quantities, namely $\mathtt{CTM}_p(1-\delta_n)$ and γ as well as the two functions φ and Ψ_p. In the following, we focus on a nonparametric estimation of φ and Ψ_p using NNs whose first step consists in approximating the two unknown functions.

Theorem 2 *Let* L *be a slowly-varying function verifying* $(\mathbf{B_J})$, $(\mathbf{C_J})$ *with* $J \geq 2$ *and* $(\mathbf{D})$*. Let* $p\gamma \in (0,1)$ *and* $\varepsilon \in (0, |\rho_2|)$*. Then, there exists a NN approximation of the CTM involving* $K_1 \geq 6$ *ReLU units and* $K_2 = J(J-1)/2$ *eLU units such that*

$$\frac{1}{p}\inf_{\theta\in\Theta}\Big|\log \mathtt{CTM}_p(1-\alpha_n) - \log \widetilde{\mathtt{CTM}}^K_{p,\theta}(1-\alpha_n; 1-\delta_n)\Big| = O(K_1^{-1\vee(\rho_2+\varepsilon)}) + O((\delta_n/\alpha_n)^{\varepsilon}\alpha_n^{-\bar\rho_J}),$$

where (α_n) *and* (δ_n) *are probability sequences such as* $\delta_n \to 0$ *and* $\alpha_n/\delta_n \to 0$ *as* $n \to \infty$*.*

The approximation rate does not depend on the tail heaviness γ, but rather on the parameters ρ_2 and $\bar\rho_J$ stemming from the J-th order condition. Since $\bar\rho_J$ is a decreasing function of J, choosing $J > 2$ may be of interest to reduce the approximation error while usual extreme-value estimators rely on first- or second-order approximations. The approximation rate is also driven by the complexity of the NN which can be quantified by both K_1 the number of ReLU units and K_2 the number of eLU units via J. Finally, note that the approximation error is an increasing function of α_n: The further in the tail the risk measure is, the better its approximation.

3 Application to the estimation of the CTM

Here, we let $\delta_n = k/n$ where $k = k_n$ is an intermediate sequence *i.e.* such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. The CTM at the intermediate level δ_n is estimated by its empirical counterpart (2) and the CTM at the extreme level α_n is thus estimated by

$$\begin{aligned}\widehat{\mathtt{CTM}}^{\mathtt{NN}}_{p,\theta}(1-\alpha_n;1-k/n) &:= \left(\frac{1}{k}\sum_{j=1}^{k} X^p_{n-j+1,n}\right)(k/(n\alpha_n))^{p\hat{\theta}_0}\exp\left(p\tilde{\varphi}^{K_2}_{\hat{\theta}_2}(\log(k/(n\alpha_n)),\log(n/k)\right)\\ &\quad\times\exp\left(\tilde{\Psi}^{K_1}_{p,\hat{\theta}_1}(\alpha_n)-\tilde{\Psi}^{K_1}_{p,\hat{\theta}_1}(k/n)\right)\\ &= \left(\frac{1}{k}\sum_{j=1}^{k} X^p_{n-j+1,n}\right)\exp\left(\tilde{g}^{K}_{p,\hat{\theta}}(\alpha_n,k/n)\right),\end{aligned}$$

where $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)$ is computed by fitting the NN model of the log-spacings to the empirical ones. More specifically, Proposition 1 shows that, for all $k \in \{1,\dots,n\}$ and $i \in \{k+1,\dots,n\}$, one has $S_{i,k} := \log \mathtt{CTM}_p(1-i/n) - \log \mathtt{CTM}_p(1-k/n) = g_p(i/n, k/n)$. The empirical estimator is deduced from (2) and given by

$$\hat{S}_{i,k} := \log\left(\frac{1}{i}\sum_{j=1}^{i} X^p_{n-j+1,n}\right) - \log\left(\frac{1}{k}\sum_{j=1}^{k} X^p_{n-j+1,n}\right),$$

while the NN model is $g_{J\phi}(i/n, k/n)$. In practice, small indices i and k are discarded since the associated empirical estimates of the CTM suffer from a large variance. To this end, we consider $\zeta \in [0,1)$ and minimize the following distance between the above two estimations of the $N = (n-2-[\zeta n])(n-1+[\zeta n]))/2$ log-spacings:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{N}\sum_{k=2+[\zeta n]}^{n-1}\sum_{i=1+[\zeta n]}^{k-1}\left|\hat{S}_{i,k} - g_{J\phi}\left(i/n,k/n\right)\right|.$$

The finite sample behavior of the estimator is illustrated in [2] both on simulated and real data sets. An overview of the results will be presented during the talk.

References

[1] M. Allouche, S. Girard, and E. Gobet. Estimation of extreme quantiles from heavy-tailed distributions with neural networks. *Statistics and Computing*, 34:12, 2024.

[2] M. Allouche, S. Girard, and E. Gobet. Learning extreme expected shortfall and conditional tail moments with neural networks. application to cryptocurrency data. *Neural Networks*, 182:106903, 2025.

[3] V. Brazauskas, B. Jones, M. Puri, and R. Zitikis. Estimating conditional tail expectation with actuarial applications in view. *Journal of Statistical Planning and Inference*, 138(11):3590–3604, 2008.

[4] E. Deme, S. Girard, and A. Guillou. Reduced-bias estimators of the conditional tail expectation for heavy-tailed distributions. In M. Hallin et al., editor, *Mathematical Statistics and Limit Theorems*, pages 105–123. Springer, 2015.

[5] J. El Methni, L. Gardes, and S. Girard. Non-parametric estimation of extreme risk measures from conditional heavy-tailed distributions. *Scandinavian Journal of Statistics*, 41(4):988–1012, 2014.

[6] I. Fraga Alves, L. de Haan, and T. Lin. Third order extended regular variation. *Publications de l'Institut Mathématique*, 80(94):109–120, 2006.

[7] Y. Goegebeur and T. de Wet. Estimation of the third-order parameter in extreme value statistics. *Test*, 21(2):330–354, 2012.

[8] M. I. Gomes, L. de Haan, and L. Peng. Semi-parametric estimation of the second order parameter in statistics of extremes. *Extremes*, 5(4):387–414, 2002.

[9] P. Hall and A. W. Welsh. Adaptive estimates of parameters of regular variation. *The Annals of Statistics*, 13:331–341, 1985.

[10] X. Q. Wang and S. H. Cheng. General regular variation of the n-th order and 2nd order Edgeworth expansions of the extreme value distribution. II. *Acta Mathematica Sinica (English Series)*, 22(1):27–40, 2006.

[11] I. Weissman. Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73(364):812–815, 1978.

CLTs for empirical processes of cluster functionals of extremes in non-mixing time series

José Gregorio Gómez-García

Université Paris Cité, CNRS, MAP5, F-75006 Paris, France.
jose-gregorio.gomez-garcia@u-paris.fr

Résumé. Comprendre comment les valeurs extrêmes se regroupent dans les séries temporelles est une question centrale en théorie des valeurs extrêmes et en gestion des risques. Drees et Rootzén (2010) ont apporté une étude approfondie à ce problème à travers les fonctionnelles de clusters d'extrêmes. Cependant, la littérature existante sur les théorèmes limites centraux fonctionnels (FCLT) pour les processus empiriques de fonctionnelles de clusters (EPCF) a, jusqu'à présent, été développée sous des conditions de mélange. Or, un simple modèle non-mélangeant ne satisfait pas ces hypothèses. Dans cet exposé, nous visons à inclure des modèles plus généraux de séries temporelles faiblement dépendantes. Pour de tels FCLT, nous assouplissons les conditions de mélange en les remplaçant par la τ-dépendance, en utilisant les propriétés de couplage de Dedecker & Prieur (2004). À titre d'exemple, nous proposons un FCLT sans conditions de mélange pour l'estimateur par blocs de l'indice extrémal. Nous présentons également une simulation illustrant la précision de notre résultat.

Mots-clés. Regroupement des valeurs extrêmes, Fonctionnelle de clusters d'extrêmes, Indice extrémal, Théorème limite central uniforme, τ-dépendance faible.

Abstract. Understanding how extreme values cluster in time series is a central question in extreme value theory and risk management. Drees and Rootzén (2010) provided an in-depth study of this problem through extreme cluster functionals. However, the existing literature on functional central limit theorems (FCLT) for empirical processes of cluster functionals (EPCF) has so far been developed under mixing conditions. Yet, a simple non-mixing model does not satisfy these assumptions. In this talk, we aim to include more general models of weakly dependent time series. For such FCLTs, we relax the mixing conditions by replacing them with τ-dependence, using the coupling properties of Dedecker & Prieur (2004). As an example, we propose an FCLT without mixing conditions for the block estimator of the extremal index. We also present a simulation illustrating the accuracy of our result.

Keywords. Clustering of extremes, Cluster functional of extremes, Extremal index, Uniform central limit theorem, τ-weak dependence.

1 Introduction

It is a well-known fact that extremes in independent sequences are isolated. However, this property does not always extend to dependent cases: in such situations, extremes occur

in clusters, as is generally the case for real phenomena. For instance, if we observe an extreme pollution day i (*e.g.*, measurements of PM2.5, PM10, ozone, and nitrogen dioxide exceeding a certain threshold u) at a specific location, it is likely that during the following days $i+1, i+2, \ldots, i+T-1$ extreme pollution will also be observed at this location; here T is, of course, a random variable. The same applies to various phenomena, such as heat waves or cold spells, etc.
The number $\mathbb{E}[T]$ is usually called *the (extremal) mean cluster size*, and it is the reciprocal value of the extremal index θ, provided that $\theta > 0$. Obviously it is important to estimate this value θ, in particular, Drees (2011) introduces the following functional estimator $(\widehat{\theta}^*_{n,t})_{0<t\leq 1}$ of θ:

$$\widehat{\theta}^*_{n,t} = \frac{\sum_{j=1}^{m_n} f_t(X_{n,(j-1)r_n+1}, \ldots, X_{n,jr_r})}{\sum_{j=1}^{m_n} g_t(X_{n,(j-1)r_n+1}, \ldots, X_{n,jr_r})}, \tag{1}$$

here $X_{n,1}, \ldots, X_{n,n}$ denotes a collection of normalised random variables, $m_n = \max\{j \in \mathbb{N} : j \leq n/r_n\}$ with $r_n \ll n$ and

$$f_t(x_1, \ldots, x_r) = \mathbb{I}_{\{\max_{1\leq i\leq r} x_i > 1-t\}}, \tag{2}$$

$$g_t(x_1, \ldots, x_r) = \sum_{i=1}^{r} \mathbb{I}_{\{x_i > 1-t\}}. \tag{3}$$

Both maps (2) and (3) are classical examples of extremes cluster functionals. They may simplify estimation procedures in extreme values theory, under weak dependence. For instance, the estimation by cluster functionals of the extremogram of Davis & Mikosch (2009) is given in (Gómez-García, 2018), the tail distributions is provided in (Drees & Rootzén, 2010), etc.

To extend those ideas, as in (Drees & Rootzén, 2010), let $(E, \mathcal{E})$ be a measurable subspace of $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ for some $d \geq 1$, such that $0 \in E$. Let $\{(X_{n,i})_{1\leq i\leq n}\}_{n\in\mathbb{N}}$ be a triangular-array of E-valued row-wise stationary normalised random variables, defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. These random variables are built from a stationary random process $(X_i)_{i\in\mathbb{Z}}$, in such a way that the normalisation $X_{n,i}$ maps all "non-extreme" values to zero. Additionally, we shall need that the sequence of conditional distributions of $X_{n,1}$ given $X_{n,1} \neq 0$, $\mathbb{P}(X_{n,1} \in \cdot \,|X_{n,1} \neq 0)$, converges weakly to some non-degenerate limit.

On the other hand, in order to define the cluster functionals, we follow the deterministic definition of Drees & Rootzén (2010). First, consider the set

$$E_\cup = \bigcup_{r=1}^{\infty} \Big\{(x_1, \ldots, x_r) : x_i \in E,\ 1 \leq i \leq r\Big\},$$

of E-valued sequences of finite length, which is equipped with the σ-field $\mathcal{E}_\cup$ induced by Borel-σ-fields on E^r, for $r \in \mathbb{N}^*$. Let $x \in E_\cup$, then we can write $x = (x_1, \ldots, x_r)$ for some

$r \in \mathbb{N}^*$. The **core** $x^c \in E_\cup$ of x is defined ([1]) by

$$x^c = \begin{cases} (x_{r_I}, x_{r_I+1}, \ldots, x_{r_S}), & \text{if } x \neq 0_r, \\ 0, & \text{otherwise,} \end{cases}$$

where 0_r denotes the null element in E^r, besides $r_I = \min\{i \in \{1, \ldots, r\} \ : \ x_i \neq 0\}$ (the first non-null value of the block x) and $r_S = \max\{i \in \{1, \ldots, r\} \ : \ x_i \neq 0\}$ (the last non-null value of the block x). A **cluster functional** is a measurable map $f : (E_\cup, \mathcal{E}_\cup) \longrightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$f(x) = f(x^c), \ \forall x \in E_\cup, \quad \text{and} \quad f(0_r) = 0, \ \forall r \geq 1. \tag{4}$$

Under the definition (4), it is easy to build a large amount of examples of cluster functionals. In particular, many useful examples admit the form:

$$f(x_1, \ldots, x_r) = \sum_{i=1}^{r} \phi(x_i), \tag{5}$$

where $\phi : E \to \mathbb{R}$ is a measurable function such that $\phi(0) = 0$. Functions ϕ are usually indicator functions (or products of another measurable function $H : E \longrightarrow \mathbb{R}$ with an indicator function). Other classic examples are either the component-wise maximum of a cluster,

$$f(x_1, \ldots, x_r) = \max_{1 \leq k \leq r} x_k, \tag{6}$$

or the functional f_t defined in (2), with components in $E = [0, \infty)$.

Let $\mathcal{F}$ be a class of cluster functionals. Define the **empirical process** $(Z_n(f))_{f \in \mathcal{F}}$ **of cluster functionals** in $\mathcal{F}$ by

$$Z_n(f) = \frac{1}{\sqrt{n v_n}} \sum_{j=1}^{m_n} (f(Y_{n,j}) - \mathbb{E} f(Y_{n,j})), \tag{7}$$

where $Y_{n,j}$ denotes the j-th block of r_n consecutive values of the n-th row stationary $\mathbb{X}_n = (X_{n,i})_{1 \leq i \leq n}$,

$$Y_{n,j} = (X_{n,i})_{(j-1)r_n + 1 \leq i \leq j r_n}, \qquad 1 \leq j \leq m_n, \tag{8}$$

and $m_n = \lceil n/r_n \rceil = \max\{j \in \mathbb{N} \ : \ j \leq n/r_n\}$ is the number of blocks of the n-th row $\mathbb{X}_n$. Besides, $v_n = \mathbb{P}(X_{n,1} \neq 0)$.
Observe that *if the blocks* $(Y_{n,j})_{1 \leq j \leq m_n, n \in \mathbb{N}}$ *are independent* and if we assume:

(C.1) $\mathbb{E}\left((f(Y_n) - \mathbb{E} f(Y_n))^2 \, \mathbb{I}_{\{|f(Y_n) - \mathbb{E} f(Y_n)| > \epsilon \sqrt{n v_n}\}} \right) = o(r_n v_n)$, for all $\epsilon > 0$, and all $f \in \mathcal{F}$;

[1] Note that the core also considers the null values that exist between the non-null values. For example: $(0,1,2,4,0,3,0,1,0,0)^c = (1,2,4,0,3,0,1)$, which is the smaller sub-block of $x = (0,1,2,4,0,3,0,1,0,0)$ which contains all non-null values as well as the null values between them.

(C.2) $\frac{1}{r_n v_n}\mathrm{Cov}\left(f(Y_n), g(Y_n)\right) \longrightarrow c(f,g)$, for all $f, g \in \mathcal{F}$;

with $r_n \ll v_n^{-1} \ll n$ and with $Y_n \overset{\mathcal{D}}{=} Y_{n,1}$ (a "generic block" because $(X_{n,i})_{1\leq i\leq n}$ is stationary for each n), then the finite-dimensional marginal distributions (fidis) of the empirical process $(Z_n(f))_{f\in\mathcal{F}}$ of cluster functionals converge to the fidis of a centred Gaussian process $(Z(f))_{f\in\mathcal{F}}$ with covariance function c.

The remainder of the paper is organized as follows. In Section 2, we briefly recall some properties related to τ-weak dependence. In Section 3, we introduce a specific family of cluster functionals, called C-Lipschitz cluster functionals, which allows us to approximate general cluster functionals by Lipschitz cluster functionals. This approximation enables the use of coupling arguments over random blocks, leading to the convergence of the finite-dimensional distributions (fidis) of the empirical process and its corollaries. The uniform convergence of the empirical process and its application to the functional block estimator of the extremal index will be presented during the talk.

2 Weak dependence

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $\mathcal{M}$ be a σ-algebra of $\mathcal{A}$ and let X be a random variable with values in a Polish space E endowed with its metric δ. Suppose that X is $\mathbb{L}^p$-integrable (*i.e.* X satisfies $\|\delta(X,0)\|_p = \left(\int \delta^p(x,0)P_X(dx)\right)^{\frac{1}{p}} < \infty$). Dedecker (2004) defined the **coefficient** τ_p as:

$$\tau_p(\mathcal{M}, X) = \left\|\sup\left\{\mathbb{E}\left[h(X)|\mathcal{M}\right] - \mathbb{E}\left[h(X)\right] \ : \ h \in \Lambda(E,\delta)\right\}\right\|_p \tag{9}$$

where $\Lambda(E,\delta)$ denotes the class of all Lipschitz functions $h : E \longrightarrow \mathbb{R}$ such that

$$\mathrm{Lip}(h) = \sup_{x\neq y} \frac{|h(x) - h(y)|}{\delta(x,y)} \leq 1.$$

Let $\mathcal{X} = \{(X_{n,i})_{1\leq i\leq n}\}_{n\in\mathbb{N}}$ be a triangular array of $\mathbb{L}^p$-integrable E-valued random variables. For each $i \in \{1,\ldots,n\}$, we set $\mathcal{M}_{n,i} = \sigma(X_{n,j},\ j \leq i)$, the σ-algebra of $\mathcal{A}$ generated by the normalised random variables $(X_{n,j})_{j\leq i}$.
Then, we define the coefficient:

$$\tau_{p,r}(k) = \max_{1\leq l\leq r} \frac{1}{l} \sup\left\{\tau_p(\mathcal{M}_{n,i}, (X_{n,j_1},\ldots,X_{n,j_l})),\ i+k \leq j_1 < \cdots < j_l \leq n\right\}, \tag{10}$$

where we consider here the metric

$$\delta_l(x,y) = \sum_{i=1}^{l} \delta(x_i, y_i) \tag{11}$$

on E^l. We then say that $\mathcal{X}$ is τ_p-weakly dependent if

$$\lim_{k\longrightarrow\infty} \limsup_{r\longrightarrow\infty} \tau_{p,r}(k) = 0. \tag{12}$$

3 CLTs for Empirical Processes of Cluster Functionals

The proof of the convergence of the fidis of the empirical process (7) still relies on the Bernstein blocks technique. Specifically, from each block $Y_{n,j}$ of length r_n, we extract a sub-block of length $l_n = o(r_n)$. The remaining sub-blocks, which are separated by l_n variables and satisfy appropriate τ-dependence conditions, enable us to couple independent blocks to the original sub-blocks, ultimately leading to the CLT.

To apply this coupling argument, we need to approximate the cluster functionals $f \in \mathcal{F}$ by a sequence of Lipschitzian cluster functionals $\{f_n\}_{n\geq 0}$. However, for this approximation to be meaningful, we must assume that the functionals $f \in \mathcal{F}$ satisfy certain properties, which we define below.

Definition 1 *Let f be a cluster functional. For each $r \in \mathbb{N}$, we denote $D^r(f)$ the set of discontinuities of $f|_{E^r}$ and $C^r(f) = E^r \setminus D^r(f)$. Let $\{\mathcal{C}_k^r\}_{k\in\Lambda(r)}$ be the distinct connected components of $C^r(f)$. We say that f is **C-Lipschitz** if for each $r \in \mathbb{N}$, $D^r(f)$ is r-null Lebesgue, $\Lambda(r)$ is finite and f is Lipschitz on each component $\mathcal{C}_k^r$.*

Remark 1 (i) It is clear that if f is Lipschitz, then it is C-Lipschitz. Therefore the functional defined in (6) is C-Lipschitz.
(ii) Let $\phi : E \longrightarrow \mathbb{R}$ be a function such that $\phi(0) = 0$ and let $D(\phi)$ be the set of discontinuities of ϕ. If $\mathrm{Card}(D(\phi))$ is finite and if ϕ is Lipschitz on each maximal connected component of $E \setminus D(\phi)$, then the functional defined in (5) is also C-Lipschitz.

On the other hand, in order to make negligible the influence of the small blocks with length l_n on the empirical process (7), we introduce assumptions over these small blocks. For this, consider the following notations.

Notation 1 Let $x = (x_1, x_2, \ldots, x_r)$. We set $x^{(l:k)}$ as follows

$$x^{(l:k)} = \begin{cases} 0, & \text{if} \quad r < l, \\ (x_l, \ldots, x_k), & \text{if} \quad 1 \leq l \leq k \leq r, \\ (x_l, \ldots, x_r), & \text{if} \quad k > r. \end{cases}$$

and $x^{(k)} = x^{(1:k)}$.
Additionally, if $f \in \mathcal{F}$ is a cluster functional, we denote

$$\Delta_n(f) = f(Y_n) - f(Y_n^{(r_n - l_n)}), \tag{13}$$

where r_n is the length of the block Y_n and $l_n < r_n$.

The following assumption guarantees that the extraction of small blocks does not interfere with the convergence of fidis of $(Z_n(f))_{f\in\mathcal{F}}$.

(C.3) For all $f \in \mathcal{F}$,

$$\begin{aligned} \mathbb{E}|\Delta_n(f) - \mathbb{E}\Delta_n(f)|^2 \mathbb{I}_{\{|\Delta_n(f)-\mathbb{E}\Delta_n(f)|\leq\sqrt{nv_n}\}} &= o(r_n v_n), \\ \mathbb{P}\left(|\Delta_n(f) - \mathbb{E}\Delta_n(f)| > \sqrt{nv_n}\right) &= o(r_n/n). \end{aligned}$$

Additionally, in order to obtain the convergence of fidis of the empirical process of cluster functionals evaluated on the big-blocks, the following assumptions are necessary:

(B) The integer sequences l_n and r_n are such that $l_n \ll r_n \ll v_n^{-1} \ll n$ and $(nv_n)^{-\gamma/2-\eta(1+\beta)} = o(r_n/n)$ for some positive constants β, γ, η, where $l_n \longrightarrow \infty$ and $nv_n \longrightarrow \infty$ as $n \to \infty$;

(C) For all $f \in \mathcal{F}$, there exist positive constants C and β such that

$$\mathbb{P}\left(Y_n \in D_{\rho_n}^{r_n}(f)\right) \leq C\rho_n^{\beta},$$

where, $D_\rho^r(f) = \bigcup_{x \in D^r(f)} \{y \in E^r : \delta_r(x,y) < \rho\}$ with $\rho > 0$;

(D.1) The triangular array $\{(X_{n,i})_{1 \leq i \leq n}\}_{n \in \mathbb{N}}$ is τ-weakly dependent such that

$$\tau_{1,r_n}(l_n) = o\left(n^{-1}(nv_n)^{\gamma/2}\right),$$

for some $\gamma \in (0,1)$;

(C.4) For all $f \in \mathcal{F}$, $\mathbb{E}\,|f(Y_n)|^{2+\alpha} = O(r_n v_n)$ and $\mathbb{P}\left(|f(Y_n)| > (nv_n)^{\frac{1-\gamma}{2}-\eta}\right) = o\left(n^{-1} r_n (nv_n)^{\frac{\alpha}{2(1+\alpha)}}\right)$ for some $\alpha > 0$, $\gamma \in (0,1)$ and $\eta \in (0, \frac{1-\gamma}{2})$.

Note that Assumption (C.4) implies Assumption (C.1).

Theorem 1 *Let $\mathcal{F}$ be a family of C-Lipschitz cluster functionals such that (C.2) and (C.3) hold. Then, if there exist α, β, γ and η such that Assumptions (B), (C), (D.1) and (C.4) are satisfied simultaneously, the fidis of the cluster functionals empirical process $(Z_n(f))_{f \in \mathcal{F}}$ converge to the fidis of a centred Gaussian process $(Z(f))_{f \in \mathcal{F}}$ with covariance function c defined in Assumption (C.2).*

Bibliographie

Basrak, B., Segers, J. (2009) *Regularly varying multivariate time series.* Stoch. Proc. Appl. 119,

Davis R. A., Mikosch, T. (2009) *The extremogram: a correlogram for extreme events.* Bernoulli 15, 977-1009.

Dedecker, J., Prieur, C. (2004) *Coupling for τ-dependent sequences and applications.* Journal of Theor. Probab. 17-4, 861-885.

Dedecker, J., Doukhan, P., Lang, G., León, J. R., Louhichi, S., Prieur, C. (2007) *Weak dependence: with examples and applications.* Lecture Notes in Statistics 190, Springer-Verlag.

Drees, H., Rootzén, H. (2010) *Limit Theorems for Empirical Processes of Cluster Functionals.* Ann. of Statist. 4, 2145-2186.

Gómez-García, J.G. (2018) *Dependent Lindeberg central limit theorem for the fidis of empirical processes of cluster functionals.* Statistics, 52(5), 955–979.

Van Der Vaart, A. W., Wellner, J. A. (1996) *Weak Convergence and Empirical Processes.* Springer, New York.

Introduction: Modelling Directional Dependence of Extreme Events*

Maxime L. D. Nicolas[1]

[1] *University College London, UK*

January 28, 2025

Abstract

This paper introduces a novel measure to quantify the directional dependence of extreme events between two variables. We propose a new approach to capture the asymmetric tail dependence. By studying the conditional tail expectation and of the rank transformed variables, we quantify the behaviour of one variable when the other is extreme. The effectiveness of the approach is demonstrated through an extensive simulation framework, and the theoretical asymptotic behaviour of the estimator is investigated. We apply this method to environmental and financial data, such as wind speed and temperature extremes, and stock-bond market extremes. Our results show the strong asymmetric nature of extreme events.

Keywords: Tail Dependence, Directional Dependence, Copula, Tail Expectation

*.
Corresponding author: M.L.D. Nicolas. E-mail address: m.nicolas@ucl.ac.uk.

1 Motivation

Extreme events are inherently multivariate across various fields such as finance, insurance, hydrology, and engineering. Typically, the probability of observing simultaneous extreme events between random variables is used to quantify extremal dependence. This is often referred to as the tail dependence coefficient and other derived measures.
However, this quantity does not provide information about the behavior of "isolated" extreme events (i.e., extreme events that do not occur together). Empirically, asymmetrical behavior is often observed. For instance, in finance, the extreme events of an asset are more likely to occur in isolation from the market rather than the opposite (individual extreme events, unexpected news, etc.). While extreme events in the market trigger extreme events in an asset most of the time.
Jasson (2005) explains asymmetric tail dependence by noting that strong storms often originate over the Atlantic and are carried from west to east by prevailing winds. For example, a severe storm affecting Germany likely already impacted France, creating a high joint probability of extreme events in both countries. However, the reverse is not necessarily true: a strong storm in France does not always continue to Germany.
Consider the case where extreme events of a variable X always occur simultaneously with those of another variable Y, while the extreme events of Y are completely independent of the occurrence of X's extreme events, and in the second case consider the same but opposite scenario for X and Y. These two distinct scenarios can lead to the same tail dependence profile if only the tail dependence coefficient is computed. Therefore, the quantity of interest should capture how "distant" or how "non-extreme", on average, thereby capturing the quantity of "isolated" extremes.
How to measure if the extreme events of a random variable X are more isolated (or independent) from those of a random variable Y? To examine tail asymmetric behaviour, researchers such as Hua and Joe (2014); Hua et al. (2019) have proposed looking at the conditional tail expectation (CTE) of the form $\mathbb{E}\left[X \mid Y > t\right]$ and $\mathbb{E}\left[X \mid Y = t\right]$. However, CTEs depend on both the dependence structures and the marginal distributions, which complicates the study. This property is also used to defined the tail exchangeable index Hua et al. (2019).

2 Directional Tail Dependence

We propose a novel measure to quantify the difference between the behaviour of X when Y is small, and the behaviour of Y when X is small.
We are therefore interested in the following quantity: the marginal expected shortfall (MSE), also known as the Conditional Tail Expectation (CTE)[1], defined as $\mathbb{E}\left[X \mid Y \leqslant t\right]$ or $\mathbb{E}\left[X \mid Y = t\right]$ for small values of t, which captures the expected behaviour of X when Y is extreme. The major drawback of this quantity is that the comparison $\mathbb{E}\left[Y \mid X \leqslant t\right]$ is not possible when X and Y have different marginal cumulative distribution functions (cdfs). Two bivariate random vectors with arbitrary marginals can only be compared if they are first transformed into the same Fréchet class Dette et al. (2013).

[1] Although they largely describe the same concept, the CTE can be defined in a univariate context. It is equivalent to the Tail Value at Risk (CVaR) and represents the expected value of X given that X exceeds its $(1-p)$-th quantile. This can be written as $\mathbb{E}\left[X \mid X > Q_X(1-p)\right]$, where Q_X is the quantile function of X.

That's why we propose applying the probability integral transform to ensure that X and Y follow standard uniform marginals using F_x and F_y, the univariate marginal cdfs of X and Y, respectively. These correspond to uniform scores and, empirically, to scaled ranks, allowing for a comparison between X and Y values in terms of their levels of extremeness.
Thus, we define the transformed conditional tail expectation:

$$\chi_{12} = \lim_{v \to 0^+} \mathbb{E}\left[F_x(X) \mid F_y(Y) \leqslant v\right]. \tag{1}$$

The CTE therefore transforms into $\mathbb{E}\left[F_x(X) \mid Y \leqslant t\right]$ and $\mathbb{E}\left[F_y(Y) \mid X \leqslant t\right]$, which translate to the expected scaled rank of a random variable when the other is small.[2]
Using the quantile function, the condition $X \leq t$ for small t is equivalent to $U_x \leq v$ for $v \to 0^+$, where $t = F_y^{-1}(v)$.
In the case of marginal independence, we have $\mathbb{E}[F_x(X) \mid Y \leq t] = \mathbb{E}[F_x(X)] = 1/2$, but the reverse does not necessarily hold.

- if $\chi_{12} < 1/2$ then positive tail influence of Y on X
- if $\chi_{12} > 1/2$ then (anti dependence) tail influence of Y on X

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be independent and identically distributed observations from the joint distribution of (X, Y). Let us define the empirical distribution function $\hat{F}_x(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}}$ where $\mathbb{1}_{\{x_i \leq x\}}$ is the indicator function that equals 1 if $x_i \leq x$ and 0 otherwise, similarly for $\hat{F}_y(y)$
The empirical version of $\chi(X, Y)$ is given by

$$\hat{\chi}_{12}\left(k/n\right) = \frac{\sum_{i=1}^{n} \hat{F}_x\left(x_i\right) \mathbb{1}\left\{\hat{F}_y\left(y_i\right) \leqslant \frac{k}{n}\right\}}{\sum_{i=1}^{n} \mathbb{1}\left\{\hat{F}_y\left(y_i\right) \leqslant \frac{k}{n}\right\}} \tag{2}$$

A proper value of $k \in (0, n]$ must be chosen, close to 0, so that $\hat{F}_x^{-1}()$ and $\hat{F}_y^{-1}()$ yield small quantiles while ensuring that there are enough values of y_i and x_i in the respective conditioning sets to compute the average.
The value of k should be selected as a function of the sample size n, such that $k \to 0$ as $n \to \infty$ to avoid bias. However, k must also be large enough to reduce the standard error of the estimator.

Definition 2.1 (Directional Tail Dependence) *Let (X, Y) be a bivariate random vector with marginal cumulative distribution functions (CDFs) F_x and F_y. Let $U_x := F(x)$ and $U_y := F(y)$ be the probability integral transforms of X and Y, respectively. The bivariate random vector (X, Y) is said to exhibit **directional tail dependence** (DTD) if the following condition is satisfied:*

$$\chi(X, Y) = \chi_{12} - \chi_{21} \neq 0. \tag{3}$$

*If condition (3) does not hold, (X, Y) are said to have **non-directed tail dependence**. The sign of $\chi(X, Y)$ indicates the direction:*

- *If $\chi(X, Y) > 0$, we say that Y is **tail directed** (TD) to X.*

[2]This quantity is also defined as the *causal tail coefficient* in Gnecco et al. (2021). However, they assume a linear structural causal model (Pearl, 2009), and it is not used to compute the strength of the asymmetric behaviour but to assert a causal link. We provide a more general definition based on asymmetric behaviour. They assume a causal link when one coefficient equals 0 (1 in their notation) and the other lies in $(0, 1/2)$.

- *If $\chi(X,Y) < 0$, we say that X is **tail directed** (TD) to Y.*

The level of $\chi(X,Y)$ describes the strength of the asymmetric (tail) behaviour between X and Y. $\chi(X,Y) \gg 0$ indicates that the values of Y are more likely to be small when X is extreme, compared to the reverse scenario. This suggests a stronger influence of X on the lower tail behaviour of Y. Conversely, $\chi(X,Y) \ll 0$ indicates that the values of X are more likely to be small when Y is extreme, compared to the reverse scenario. This suggests a stronger influence of Y on the lower tail behaviour of X.
Although necessary requirements for a directed dependence metric have been previously specified (Junker et al., 2021), the coefficient $\chi(X,Y)$, by definition, follows distinct and additional properties:

1. $\chi(X,Y) \in [-1,1]$ (*boundedness*)
2. $\chi(X,Y) = -\chi(Y,X)$ (*antisymmetry*)
3. $X \perp\!\!\!\perp Y \Rightarrow \chi(X,Y) = \chi(Y,X) = 0$ (*nullity when independence*)
4. $\lambda_{XY} = 1 \Rightarrow \chi(X,Y) = \chi(Y,X) = 0$ (*nullity when complete tail dependence*)
5. $(X,Y) \overset{d}{=} (Y,X) \Rightarrow \chi(X,Y) = \chi(Y,X) = 0$ (*exchangeability*)
6. $\chi\big(g(X), h(Y)\big) = \chi(X,Y)$, for $g, h : \mathbb{R} \to \mathbb{R}$ strictly monotonic functions (*scale invariance*)

Remark 1. Exchangeability of the dependence structure of two random variables X and Y means that the joint distribution of X and Y is invariant under permutation. Thus, $(X,Y) \overset{d}{=} (Y,X)$, where $\overset{d}{=}$ denotes equality in distribution. This implies that the joint probability distribution satisfies: $P(X \le x, Y \le y) = P(Y \le x, X \le y) \quad$ for all $x, y \in \mathbb{R}$. Although it is evident that the coefficient $\chi(X,Y)$ is a property of the copula, it differs from the previously defined *tail exchangeability* Hua et al. (2019), which is defined for the survival function of non-negative random variables only.
Remark 2. Having $\chi(X,Y) \to 0$ does not provide information about the strength of extremal dependence between X and Y. For example, even in cases of perfect extremal dependence ($\lambda_{XY} = 1$), we may observe a null directed tail dependence (DTD), similar to when X and Y are independent.
Remark 3. Having $\big|\chi(X,Y)\big| > 1/2$ indicates anti-dependence. Furthermore, $\big|\chi(X,Y)\big| \to 1$ if and only if there is perfect (complete) extremal anti-dependence, meaning one variable is left-tail extreme only when the other is right-tail extreme.

Proposition 2.1 *We can express the transformed CTE in terms of (conditional) copula C.*

$$\chi_{12} = \frac{1}{v}\int_0^1 u \frac{\partial C(u,v)}{\partial u} du, \quad v \longrightarrow 0^+. \tag{4}$$

Using the monotone convergence theorem, we can write:

$$\chi_{12} = \int_0^1 u \Delta_{2|1}(u) du, \tag{5}$$

where $\Delta_{2|1}(u) = \lim_{v \to 0} \frac{C_{2|1}(v|u)}{v}$.

References

Dette, H., Siburg, K. F., and Stoimenov, P. A. (2013). A copula-based non-parametric measure of regression dependence. *Scandinavian Journal of Statistics*, 40(1):21–41.

Gnecco, N., Meinshausen, N., Peters, J., and Engelke, S. (2021). Causal discovery in heavy-tailed models. *The Annals of Statistics*, 49(3):1755–1778.

Hua, L. and Joe, H. (2014). Strength of tail dependence based on conditional tail expectation. *Journal of Multivariate Analysis*, 123:143–159.

Hua, L., Polansky, A., and Pramanik, P. (2019). Assessing bivariate tail non-exchangeable dependence. *Statistics & Probability Letters*, 155:108556.

Jasson, S. (2005). L'asymétrie de la dépendance, quel impact sur la tarification. In *Technical Report*. AXA Group Risk Management Paris, France.

Junker, R. R., Griessenberger, F., and Trutschnig, W. (2021). Estimating scale-invariant directed dependence of bivariate distributions. *Computational Statistics & Data Analysis*, 153:107058.

Pearl, J. (2009). *Causality*. Cambridge university press.

Penalized estimation of GEV parameters for extreme quantile regression

Lucien M. Vidagbandji[1] & Alexandre Berred [1] & Cyrille Bertelle [2] & Laurent Amanton [2]

[1] *Université Le Havre Normandie, Normandie Univ., LMAH UR 3821.*
mahutin-lucien.vidagbandji@univ-lehavre.fr ; alexandre.berred@univ-lehavre.fr
[2] *Université Le Havre Normandie, Normandie Univ., LITIS, France.*
cyrille.bertelle@univ-lehavre.fr ; laurent.amanton@univ-lehavre.fr

Résumé. La régression quantile (QR) repose sur l'estimation des quantiles conditionnels et permet d'explorer les relations entre les variables indépendantes et dépendantes. Cependant, elle rencontre des difficultés d'extrapolation pour des niveaux de probabilité élevés en raison de la rareté des données dans les queues de distribution, ainsi que des défis liés à la complexité des fonctions de quantile et au nombre potentiellement élevé de prédicteurs. Pour résoudre ces problèmes, nous proposons une méthode qui modélise les maxima par blocs à l'aide de la distribution des valeurs extrêmes généralisée conditionnelle (GEV), dont les paramètres dépendent des covariables, ce qui améliore l'extrapolation dans les queues. Nous utilisons ensuite une approche basée sur les forêts aléatoires généralisées (GRF) pour estimer ces paramètres. Concrètement, nous maximisons une vraisemblance pénalisée pondérée par les poids issus de la méthode GRF. Cette pénalisation permet de surmonter les limites de l'estimateur du maximum de vraisemblance (MLE) pour les petits échantillons tout en maintenant son optimalité pour les grands échantillons. La performance de notre méthode est illustrée par des études de simulation et une application aux données de salaires aux États-Unis de 1980.

Mots-clés. Régression quantile, distribution des valeurs extrême généralisée, forêt aléatoire généralisée, maximum de vraisemblance, bloc maxima, quantile extrême.

Abstract. Quantile regression (QR) relies on the estimation of conditional quantiles and allows for the exploration of relationships between independent and dependent variables. However, it faces extrapolation difficulties at high probability levels due to the scarcity of data in the tails of the distribution, as well as challenges related to the complexity of quantile functions and the potentially large number of predictors. To address these issues, we propose a method that models block maxima using the conditional generalized extreme value (GEV) distribution, whose parameters depend on covariates, thus enhancing extrapolation in the tails. We then use an approach based on generalized random forests (GRF) to estimate these parameters. Specifically, we maximize a penalized likelihood weighted by the weights obtained from the GRF method. This penalization overcomes the limitations of the maximum likelihood estimator (MLE) for small samples while maintaining its optimality for large samples. The performance of our method is demonstrated through simulation studies and an application to U.S. wage data from 1980.

Keywords. Quantile regression, generalized extreme value distribution, generalized random forest, maximum likelihood, block maxima, extreme quantile.

1 Introduction

The analysis of extreme phenomena is crucial due to their growing impacts, particularly in the fields of climate change, financial crises, and technological failures. Although rare, these events have severe and disproportionate consequences, occurring in the extreme tails of statistical distributions, which necessitates the use of specialized modeling tools. Introduced by Koenker and Bassett (1978), the quantile regression model allows for the estimation of a conditional quantile from a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, representing independent copies of the random vector (X, Y), where $X \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathcal{Y} \subset \mathbb{R}$. This model estimates the conditional quantile :

$$\mathcal{Q}_x(\tau) = F^{-1}_{Y|X=x}(\tau), \tag{1}$$

where $x \in \mathcal{X}$ and $\tau \in (0, 1)$. When $\tau \to 1$, estimating this quantile requires extrapolation into the tail of the distribution, which is challenging due to the scarcity of data in these regions. Extreme quantile regression, leveraging extreme value theory, addresses this challenge. Several studies have integrated this theory within the framework of quantile regression, as detailed by Chernozhukov et al. (2017). Another challenge arises when the quantile function is complex and the dimension p of the predictor space is high. To address these challenges, models combining extreme value theory and statistical learning methods have recently been developed, notably by Velthoen et al. (2023), Pasche and Engelke (2023), and Gnecco et al. (2024). In this work, we propose a method for estimating the conditional quantile to tackle these challenges. To improve extrapolation in the tails of the distribution, we model the block maxima distribution using the conditional generalized extreme value (GEV) distribution, whose parameters depend on the covariates x. To address the complexity of quantile functions, high-dimensional predictors, and the limitations of the maximum likelihood estimator (MLE) on small samples, we introduce a weighted version of the penalized likelihood estimator proposed by Coles and Dixon (1999). The weights are obtained using generalized random forests (GRF) introduced by Athey et al. (2019). Simulation studies and an application to U.S. wage data demonstrate that our method outperforms those of Meinshausen and Ridgeway (2006) and Athey et al. (2019) in terms of performance.

2 Background

In the absence of covariates, consider a sample $(Y_1, \ldots, Y_n)$ of i.i.d. realizations of a random variable Y with distribution function F. The goal is to estimate the quantile $\mathcal{Q}(\tau) = F^{-1}(\tau)$ for an extreme probability level $\tau = \tau_n$ such that $\tau_n \to 1$ and $n(1-\tau_n) \to d \geq 0$ as $n \to +\infty$. In this case, the number of observations exceeding $\mathcal{Q}(\tau_n)$ becomes limited, making empirical estimation challenging and requiring extrapolation into the tail (Haan and Ferreira (2006)). An important result for extrapolation is the asymptotic theorem by Fisher and Tippett (1928) and Gnedenko (1943), which establishes the limiting distribution of the maximum of a sequence of i.i.d. variables, under the existence of sequences $a_n > 0$ and $b_n \in \mathbb{R}$ such that :

$$F^n(b_n + a_n x) \to G(x), \quad as \quad n \to \infty, \tag{2}$$

with G being a non-degenerate distribution, G belongs to one of the extreme value distributions defined by : $G_\xi(x) = \exp\left(-(1+\xi x)^{-\frac{1}{\xi}}\right)$, for all x such that $1 + \xi x > 0$ and $\xi \in \mathbb{R}$.

A distribution function F is said to belong to the domain of attraction of an extreme value distribution, denoted by $F \in \mathcal{D}(G_\xi)$, if there exist sequences $a_n > 0$ and $b_n \in \mathbb{R}$ such that the equality (2) holds. The extreme value index ξ, which indicates the heaviness of the tail, is crucial for estimating extreme quantiles. In this work, we use this distribution to facilitate the extrapolation.

3 Method

Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be a sequence of independent and identically distributed random vectors, originating from the random vector (X, Y). Suppose that the conditional cumulative distribution function of Y given $X = x$, denoted $F_x(\cdot)$, belongs to the domain of attraction of the extreme value distribution, i.e., $F_x \in \mathcal{D}\left(G_{\xi(x)}\right)$, with the corresponding normalization sequences $a_m(x)$ and $b_m(x)$ defined in (2). We divide the considered sequence of vectors into n blocks of size m, such that the k-th block, for $k \geq 1$, is given by : $B_{k,m} = \{(X_{(k-1)m+1}, Y_{(k-1)m+1}), \ldots, (X_{km}, Y_{km})\}$. Let $Z_{k,m} = \max\{Y_i : (X_i, Y_i) \in B_{k,m}\}$ and $X_{k,m}$ be the X_i associated with the Y_i that maximizes $\{Y_i : (X_i, Y_i) \in B_{k,m}\}$. Thus, for all $m > 1$, the sequence $\{(X_{k,m}, Z_{k,m})\}_{k\geq 1}$ represents the block maxima, where the maximum is taken with respect to the Y-component. For every $x \in \mathcal{X}$ and $m \geq 1$, the variables $\{Z_{k,m}|X_{k,m} = x\}_{k\geq 1}$ are independent and identically distributed with a cumulative distribution function F_x^m. Therefore, for every $x \in \mathcal{X}$, the distribution of $Z_{k,m}|X_{k,m} = x$ is given by the conditional GEV distribution, where the parameters depend on the covariate x, and it is given by :

$$G_{\theta(x)}(z) = \exp\left[-\left(1 + \xi(x)\frac{z - \mu(x)}{\sigma(x)}\right)_+^{-\frac{1}{\xi(x)}}\right], \tag{3}$$

with $\theta(x) = (\mu(x), \sigma(x), \xi(x))$ for all $x \in \mathcal{X}$. The τ-th order quantile ($\tau > \tau_0$) of the conditional GEV distribution is obtained by inverting the distribution given in (3). For all $x \in \mathcal{X}$, it is given by :

$$Q_x(\tau) = \begin{cases} \mu(x) + \dfrac{\sigma(x)}{\xi(x)}\left((-\ln(\tau))^{-\xi(x)} - 1\right) & \text{if } \xi(x) \neq 0, \\ \mu(x) - \sigma(x)\ln\left(-\ln(\tau)\right) & \text{if } \xi(x) = 0. \end{cases} \tag{4}$$

To estimate the parameters $\mu(x)$, $\sigma(x)$, and $\xi(x)$, maximum likelihood estimation (MLE) is commonly used for its flexibility and efficiency in modeling extremes. However, it performs less well than the alternative method based on probability-weighted moments (PWM) when applied to small samples (Coles and Dixon (1999)). In contrast, the PWM estimator limits the variation of ξ by imposing $\xi < 1$. To incorporate this information into the MLE estimator, Coles and Dixon (1999) propose penalizing the likelihood in the form $L_{pen}(\mu, \sigma, \xi) = L(\mu, \sigma, \xi)P_{\alpha,\lambda}(\xi)$, where $L(\mu, \sigma, \xi)$ is the GEV likelihood function.

$$P_{\alpha,\lambda}(\xi) = \begin{cases} 1 & \text{if } \xi \leq 0, \\ \exp\left\{-\lambda(\frac{1}{1-\xi} - 1)^\alpha\right\} & \text{if } 0 < \xi < 1, \\ 0 & \text{if } \xi \geq 1, \end{cases} \tag{5}$$

where $\alpha \geq 0$ and $\lambda \geq 0$ are tuning parameters. Although, to a large extent, the approach helps reduce computational burden, it retains the asymptotic properties of the classical MLE estimator. They show that the penalized maximum likelihood estimator performs similarly to the PWM estimator for small samples. We apply this method to estimate the parameters of the conditional GEV distribution in this work. The parameter estimator $\hat{\theta}(x)$ is defined as the one that minimizes the weighted and penalized negative log-likelihood, given by :

$$L_n^{pen}(\theta; x) = \sum_{i=1}^{n} w_i(x)\ell_\theta(z_i) + \log(P_{\alpha,\lambda}(\xi)), \quad \forall x \in \mathcal{X}, \tag{6}$$

where $\ell_\theta(z_i) = -\log\left(\frac{dG(z_i;\mu,\sigma,\xi)}{dz_i}\right)$ and $P_{\alpha,\lambda}(\xi)$ is the penalty function defined in (5). The optimal values of λ and α are obtained through cross-validation. The weights $w_i(x)$ for $i = 1, \ldots, n$ are obtained using the GRF method from Athey et al. (2019). The theoretical guarantees of the maximum likelihood estimator of the GEV distribution are well documented in the literature. The likelihood function of the GEV distribution does not have a global maximum, but for simplicity, we will refer to $(\hat{\mu}(x), \hat{\sigma}(x), \hat{\xi}(x))$ as a weighted penalized maximum likelihood estimator if $L_n^{\text{pen}}(\theta; x)$ has a local minimum at $(\hat{\mu}(x), \hat{\sigma}(x), \hat{\xi}(x))$. The parameter space of the GEV distribution, denoted $\theta(\mathcal{X})$, is unknown in practice. Therefore, we define our estimator $\hat{\theta}(x)$ as the value of the parameter θ that optimizes (6) over a compact set $\Theta \subset \mathbb{R} \times (0, +\infty) \times (-1, +\infty)$ such that $\theta(\mathcal{X}) \subset \Theta$. It is thus defined as : $\hat{\theta}(x) = \arg\min_{\theta \in \Theta} L_n^{pen}(\theta; x)$. These estimated parameters are substituted into (4) to give the final estimate of the conditional quantile of order $\tau > \tau_0$.

4 simulation studies

We assume that $Y|X = x \sim \gamma(x)\mathcal{T}_{\nu(x)}$, where : $\gamma(x) = 1 + 2\pi\varphi(2x_1, 2x_2)$ and $\nu(x) = 3 + \frac{7}{1+\exp(4x_1+1.2)}$, with φ representing the density of the centered bivariate normal distribution with unit variance and a correlation coefficient of 0.75. In this study, we consider complex forms for the functions $\gamma(x)$ and $\nu(x)$, which depend on the covariates x. This allows us to assess the robustness of our method against combined challenges, such as those mentioned in the introduction. The performance of our method, denoted erf_Pen, is better compared to the methods grf by Athey et al. (2019) and qrf by Meinshausen and Ridgeway (2006), as shown in figure 1.

5 Real dataset

We apply our method to the U.S. wage data from 1980 (Angrist et al. (2006)). Figure 2 shows that $\hat{\mu}(x)$ and $\hat{\sigma}(x)$ increase with the level of education, while $\hat{\xi}(x)$ decreases. The values of $\hat{\xi}(x)$, ranging from 0.1 to 0.25, indicate heavy tails, which confirms the results obtained by Gnecco et al. (2024) on the same data. Figure 3 demonstrates that our method better captures the complex structure of the data compared to the grf and qrf methods.

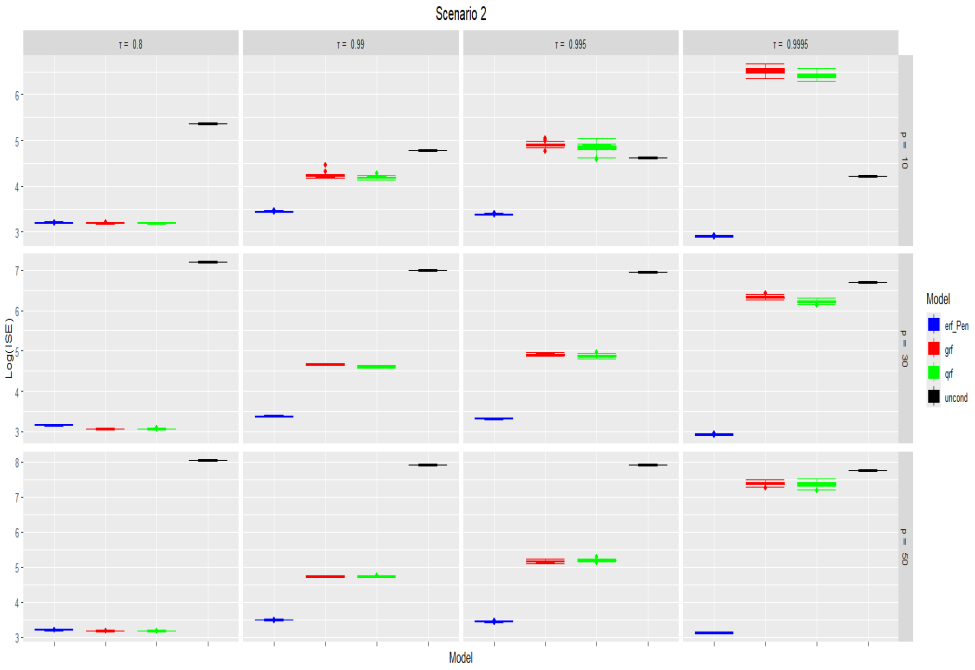


FIGURE 1 – Boxplots of Log(ISE) over 100 simulations for different values of p and extreme probability levels.

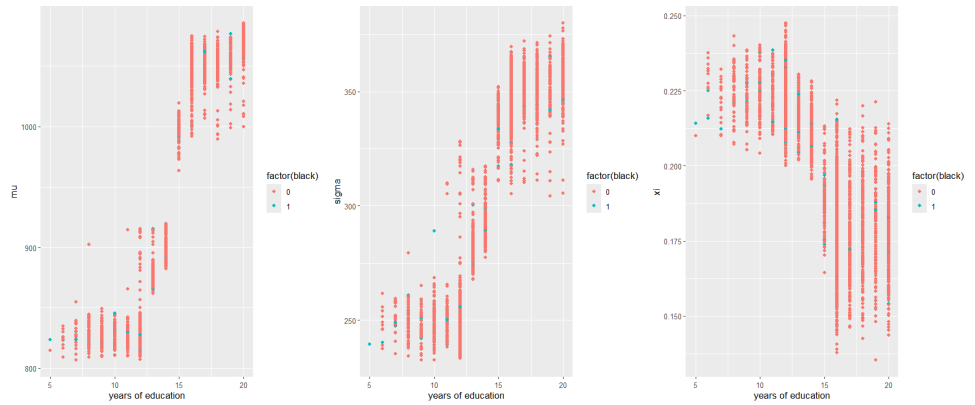


FIGURE 2 – Variations of the parameters $\hat{\mu}(x)$, $\hat{\sigma}(x)$, and $\hat{\xi}(x)$ as a function of the number of years of education.

Références

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the US wage structure. *Econometrica*, 74(2) :539–563. Publisher : Wiley Online Library.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2) :1148–1178.

FIGURE 3 – Predicted conditionnal quantiles at level $\tau = 0.8, 0.9, 0.995$ as fonction of years of eduction for erf_Pen, grf and qrf method.

Chernozhukov, V., Fernández-Val, I., and Kaji, T. (2017). *Extremal Quantile Regression : An Overview*. arXiv :1612.06850 [econ, stat].

Coles, S. G. and Dixon, M. J. (1999). Likelihood-Based Inference for Extreme Value Models. *Extremes*, 2(1) :5–23.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2) :180–190. Publisher : Cambridge University Press.

Gnecco, N., Terefe, E. M., and Engelke, S. (2024). Extremal Random Forests. *Journal of the American Statistical Association*, pages 1–24.

Gnedenko, B. (1943). Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire. *Annals of Mathematics*, 44(3) :423–453. Publisher : Annals of Mathematics.

Haan, L. and Ferreira, A. (2006). *Extreme value theory : an introduction*, volume 3. Springer.

Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica*, 46(1) :33.

Meinshausen, N. and Ridgeway, G. (2006). Quantile regression forests. *Journal of machine learning research*, 7(6).

Pasche, O. C. and Engelke, S. (2023). Neural Networks for Extreme Quantile Regression with an Application to Forecasting of Flood Risk. arXiv :2208.07590 [stat] version : 2.

Velthoen, J., Dombry, C., Cai, J.-J., and Engelke, S. (2023). Gradient boosting for extreme quantile regression. *Extremes*.

Finance, assurance

Approximate Bayesian Calibration of Option Pricing Models

Éléonore Blanchard [1] & Pierre-Olivier Goffard [2] & Patrick Laub [3] & Peter Tankov [4]

[1] *Université d'Aix-Marseille, France, eleonoredblanchard@gmail.com*
[2] *Université de Strasbourg, Institut de Recherche Mathématiques Avancées, France, goffard@unistra.fr*
[3] *UNSW, Australia, p.laub@unsw.edu.au*
[4] *ENSAE, France, peter.tankov@ensae.fr*

Résumé. Nous proposons une méthodologie pour la calibration de modèles financiers complexes basée sur l'algorithme Approximate Bayesian Computation (ABC). Cette approche permet de calibrer les paramètres de modèles nécessitant un pricing par Monte-Carlo. Nous appliquons la méthodologie à trois modèles : un modèle exponentiel de Levy, un modèle exponentiel de Levy modulé par une chaîne de Markov et un modèle de volatilité rugueuse. Nos expériences, réalisées sur données simulées et réelles (options européennes sur le S&P500), montrent que l'algorithme ABC fournit des calibrations précises, particulièrement pour le modèle Rough Bergomi où les méthodes classiques ne sont pas applicables. En complément, une approche par processus gaussiens améliore la stabilité des calibrations dans le temps. Cette étude démontre l'efficacité et la flexibilité d'ABC pour la calibration de modèles complexes, ouvrant des perspectives pour son application à des produits dérivés plus sophistiqués.

Mots-clés. Modèles de pricing, Calibration, Approximate Bayesian Computation (ABC), Statistiques Bayésiennes, Modèle exponentiel de Lévy, Diffusion à saut de Merton, Modèle de Markov à changement de régime, Volatilité rugueuse, Rough Bergomi, Pricessus gaussiens

Abstract. We present a methodology for the calibration of complex financial models based on the Approximate Bayesian Computation (ABC) algorithm. This approach enables the calibration of parameters in models requiring pricing by Monte Carlo. We apply the methodology to three models: an exponential Levy model, an exponential Levy model modulated by a Markov chain and a rough volatility model. Our experiments, conducted on simulated and real data (European options on the S&P500), show that the ABC algorithm provides accurate calibrations, particularly for the Rough Bergomi model where classical methods do not apply. Additionally, a Gaussian process-based approach improves the stability of calibrations over time. This study demonstrates the efficiency and flexibility of ABC for the calibration of complex models, opening perspectives for its application to more sophisticated derivative products.

Keywords. Pricing Models, Calibration, Approximate Bayesian Computation (ABC), Bayesian Statistics, Exponential Lévy Model, Merton Jump Diffusion, Markov-Switching Diffusion Model, Rough Volatility, Rough Bergomi, Gaussian Processes

1 Introduction

La calibration des modèles de pricing d'options consiste à ajuster leurs paramètres pour correspondre aux prix observés sur le marché, comme le décrivent Avellaneda [1]. Cette démarche commence par la sélection d'un modèle décrivant la valeur de l'actif sous-jacent $(S_t)_{t\geq 0}$. Nous travaillons sur les modèles exponentiels de Lévy, tels que discutés par Cont and Tankov [4, 5] ainsi que le modèle à volatilité rugueuse de Rough Bergomi présenté dans Bayer et al. [2].

Les données nécessaires proviennent des options européennes d'achat. Leur prix correspond à l'espérance actualisée du gain à l'échéance, faisant de la calibration un problème de moments généralisé pour le processus stochastique $(S_t)_{t\geq 0}$, comme expliqué par Hansen [8]. Sous une mesure neutre au risque, $(e^{-rt}S_t)_{t\geq 0}$ devient une martingale, où $r > 0$ est le taux d'actualisation.

Cependant, ce problème inverse est mal posé : il peut ne pas avoir de solution ou en posséder une infinité. Une approche consiste à supposer que $(S_t)_{t\geq 0}$ suit un modèle exponentiel de Lévy paramétré par $\theta \in \Theta \subset R^d$, comme dans les travaux de Yang et al. [13], Yang and Lee [12] et Guillaume and Schoutens [7]. Si les prix observés ne peuvent être reproduits exactement, on peut minimiser une mesure de l'écart, souvent une erreur quadratique moyenne pondérée. En cas de solutions multiples, une régularisation est ajoutée, impliquant une distance à un modèle a priori, comme discuté par Avellaneda [1], Cont and Tankov [4], et Cont and Tankov [5].

Pour les modèles de Lévy généraux, les prix des options ne sont pas disponibles sous forme analytique et nécessitent une inversion numérique de la transformée de Fourier, selon la méthode de Carr and Madan [3]. Ainsi, des algorithmes d'optimisation sont indispensables pour explorer l'espace des paramètres Θ. Le premier modèle considéré est le modèle de diffusion à saut proposé par Merton [10], pour lequel nous comparons la calibration avec des prix obtenus par inversion de Fourier et par des simulations de Monte-Carlo. Le deuxième modèle est le modèle de diffusion à changements de régimes de Markov, pour lequel la fonction caractéristique pour l'inversion de Fourier est plus complexe à calculer. Toutefois, une méthode alternative, plus efficace et basée sur les simulations de Monte-Carlo, est disponible : la méthode de semi Monte-Carlo. Nous comparons les résultats de calibration pour ce modèle en utilisant des prix calculés à la fois par les méthodes de Monte-Carlo classique et de semi Monte-Carlo.

Dans le modèle de Rough Bergomi, présenté dans Bayer et al. [2], la fonction caractéristique n'est pas disponible en raison de la structure particulière du processus sous-jacent, qui rend impossible l'obtention d'une expression analytique fermée. La fonction caractéristique est définie comme l'espérance $E[e^{iuX_T}]$ pour un actif X_T à une maturité T. Cependant, dans le modèle de Rough Bergomi, les dépendances à mémoire longue induites par le mouvement brownien fractionnaire empêchent la factorisation des termes pour arriver

à une expression fermée. Ainsi il n'est pas possible d'effectuer une inversion numérique de la transformée de Fourier, et seules les simulations de Monte-Carlo permettent de calculer les prix des options.

2 Problème de calibration

Soit $(S_t)_{t\geq 0}$ un processus stochastique défini sur un espace de probabilité filtré $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, P)$ qui correspond au prix d'un certain actif financier jusqu'à l'horizon T. Notre objectif est de calibrer un modèle pour la dynamique de $(S_t)_{t\geq 0}$ sur la base de données constituées de prix d'options. Une option est un contrat qui donne à l'acheteur le droit, mais non l'obligation, d'acheter ou de vendre un actif sous-jacent à un prix spécifié K dans une période de temps spécifiée délimitée par l'échéance T. Les options les plus couramment utilisées sont les options d'achat de type européen, dont le prix est donné par la formule suivante

$$C(K,T) = E^Q\left[e^{-rT}(S_T - K)_+\right]$$

où $r > 0$ représente le taux d'actualisation, généralement le taux sans risque. Les prix sont calculés en fonction de la mesure de probabilité neutre à l'égard du risque Q, équivalente à P, en vertu de laquelle les prix actualisés $(e^{-rt}S_t)_{t\geq 0}$ sont des martingales.

On suppose que nous observons les prix de plusieurs options d'achat cotées à l'instant $t = 0$, donnés par

$$C_{1:n} = \{C(T_i, K_i)\}_{i=1,\ldots,n},$$

alors le problème de calibration peut être formulé comme :

Problème 1 *Soit Q_θ une mesure de probabilité risque neutre, trouver θ pour minimiser*

$$d(C_{1:n}, C^\theta_{1:n}) = \sqrt{\sum_{i=1}^{n} w_i \left(C_i - C_i^\theta\right)^2},$$

où $w_i > 0$, $\sum_{i=1}^n w_i = 1$, et

$$C_i^\theta = C(K_i, T_i) = E^{Q\theta}\left[e^{-rT_i}(S_{T_i} - K_i)_+\right], \text{ pour } i = 1, \ldots, n.$$

3 Méthodes

Afin d'effectuer la calibration des modèles précédement décrits nous employons l'algorithme Approximate Bayesian Computation (ABC), tel que décrit dans Goffard and Laub [6]. Il est conçu pour les situations où les prix des options générés par le modèle sont calculés via des simulations Monte Carlo. Bien que ces simulations soient lentes et imprécises, elles

restent populaires pour leur flexibilité et leur simplicité. L'algorithme ABC prend en compte l'erreur d'estimation des prix générés par le modèle, reflétant cette incertitude dans les valeurs paramétriques estimées. Cela entraîne un compromis entre le nombre de répliques Monte Carlo, le temps de calcul, la précision et l'incertitude sur les paramètres.

Cette méthode consiste à traiter le paramètre comme une variable aléatoire avec une distribution a priori $p(\theta)$ que nous mettons à jour à l'aide des données du marché $C_{1:n} = \{C_1, ...C_n\}$ afin d'obtenir la distribution a posteriori $p(\theta|C_{1:n})$. Les techniques bayésiennes classiques ne conviennent pas ici, car nous ne pouvons pas écrire la fonction de vraisemblance de nos données. Nous contournons ce problème en utilisant un algorithme résumé ci-dessous.

1. Échantillonner une valeur de paramètre θ.
2. Générer les données de marché synthétiques $C^{\theta}_{1:n}$ en fonction du paramètre échantillonné θ.
3. Si les données de marché observées et synthétiques sont suffisamment proches, enregistrer la valeur du paramètre θ.

La procédure est décrite plus en détail dans **Algorithm 1**.

Algorithm 1 Population Monte Carlo Approximate Bayesian Computation

set $\epsilon_0 = \infty$ and $p_{\epsilon_0}(\mathcal{D}) = \pi()$
for $g = 1 \to G$ **do**
  for $k = 1 \to K$ **do**
    repeat
      generate $^* \sim p_{\epsilon_{g-1}}()$
      compute $C_i^* = C(K_i, T_i) = E^{Q*}\left[e^{-rT_i}(S_{T_i} - K_i)_+\right]$, for $i = 1, \ldots, n$
    until $\mathrm{d}\left(C_{1:n}, C^*_{1:n}\right) < \epsilon_g$
    set $^g_k =^*$ and $d^g_k = d^*$
  end for
  find $\epsilon_g \leq \epsilon_{g-1}$ so that$\hat{}= \left[\sum_{k=1}^K (w^g_k)^2\right]^{-1} \approx K/2$, where

$$w^g_k \propto \frac{p(^g_k)}{p_{\epsilon_{g-1}}(^g_j\mathcal{D})}_{d_k < \epsilon_g}, \quad j = 1, \ldots, N$$

  compute $p_{\epsilon_g}(\mathcal{D}) = \sum_{k=1}^K w^g_k K(-^g_k)$
end for

Les algorithmes ABC sont décrits dans le livre de Sisson et al. [11] et ont trouvé des applications réussies dans divers problèmes de finance et de gestion des risques. Pour une revue récente sur l'utilisation des méthodes bayésiennes en finance et en économie, voir Martin et al. [9].

Cet algorithme est de plus idéal afin de calibrer les paramètres des modèles pour lesquels les prix d'options peuvent uniquement être calculés par simulations de Monte-Carlo, comme c'est le cas pour le modèle de Rough Bergomi, mais également pour les options plus complexes dont le prix est dépendant de l'évolution du prix du sous-jacent tout au long de la durée de vie de l'option (*path dependant*), comme les options asiatiques, barrières ou encore les produits exotiques.

La performance de l'algorithme est influencée par différents paramètres clés : le nombre de générations G ou la valeur minimale ϵ_g acceptée, la taille de la population K, le nombre de réplications Monte Carlo utilisées pour calculer le prix des options, et le nombre de prix d'options calculés n. Comme on peut s'y attendre, le temps de calcul de l'algorithme augmente avec les valeurs élevées de ces paramètres. Il faut donc faire un arbitrage afin de trouver les valeurs optimales des paramètres.

Dans le cadre de cette présentation nous nous concentrerons uniquement sur le cas des options d'achat vanilles, mais ces autres applications sont tout à fait envisageables dans le cadre d'une autre étude pour illustrer notre méthodologie.

4 Applications

Nous appliquons cette calibration dans une étude de simulations afin de comparer les valeurs obtenues avec la calibration et les vrais valeurs de paramètres qui ont permis de générer les observations synthétiquement. Elle nous permet également de choisir les paramètres optimaux de l'algorithme.

Nous appliquons ensuite notre méthode sur des prix d'options d'achats sur le S&P500 observés de janvier 2020 à décembre 2024 pour différentes maturités et prix d'exercice. Nous introduisons une approche basée sur des processus gaussiens afin de lisser les calibrations dans le temps, réduisant ainsi la variabilité des paramètres estimés. Cette étude nous permet d'observer l'évolution des paramètres dans le temps et d'observer des évolutions sur le marché, notamment en terme de volatilité implicite.

Nos résultats montrent que l'utilisation d'ABC apporte une contribution scientifique significative, en particulier pour le modèle Rough Bergomi où les approches traditionnelles sont limitées. De plus, la méthode semi-Monte Carlo offre une alternative rapide et efficace pour le modèle Markov-Switching, tout en maintenant une précision satisfaisante. Enfin, l'approche par processus gaussiens améliore la robustesse des calibrations sur données réelles.

5 Conclusion

En conclusion, cette étude démontre l'efficacité de l'algorithme ABC pour la calibration de modèles financiers nécessitant un pricing avec Monte-Carlo. Elle ouvre également des perspectives pour l'utilisation d'approches similaires dans la calibration de produits dérivés plus complexes. La généralisation de la méthode à tout processus pouvant être simulé et évalué par Monte-Carlo est une contribution intéressante.

Bibliographie

[1] Marco Avellaneda. Minimum-relative-entropy calibration of asset-pricing models. *International Journal of Theoretical and Applied Finance*, 01(04):447–472, 1998. doi: 10.1142/S0219024998000242. URL https://doi.org/10.1142/S0219024998000242.

[2] Christian Bayer, Peter K. Friz, and Jim Gatheral. Pricing under rough volatility. *SSRN Electronic Journal*, 2015. ISSN 1556-5068. doi: 10.2139/ssrn.2554754.

[3] Peter Carr and Dilip Madan. Option valuation using the fast fourier transform. *The Journal of Computational Finance*, 2(4):61–73, 1999. ISSN 1460-1559. doi: 10.21314/jcf.1999.043.

[4] Rama Cont and Peter Tankov. Nonparametric calibration of jump-diffusion option pricing models. *The Journal of Computational Finance*, 7:1–49, 2004. URL https://hal.science/hal-00002694. 2003.

[5] Rama Cont and Peter Tankov. Retrieving lévy processes from option prices: Regularization of an ill-posed inverse problem. *SIAM Journal on Control and Optimization*, 45(1):1–25, January 2006. ISSN 1095-7138. doi: 10.1137/040616267.

[6] Pierre-Olivier Goffard and Patrick J. Laub. Approximate bayesian computations to fit and compare insurance loss models, 2020.

[7] Florence Guillaume and Wim Schoutens. A moment matching market implied calibration. *Quantitative Finance*, 13(9):1359–1373, September 2013. ISSN 1469-7696. doi: 10.1080/14697688.2013.794950.

[8] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029, July 1982. ISSN 0012-9682. doi: 10.2307/1912775.

[9] Gael M. Martin, David T. Frazier, Worapree Maneesoonthorn, Rubén Loaiza-Maya, Florian Huber, Gary Koop, John Maheu, Didier Nibbering, and Anastasios Panagiotelis. Bayesian forecasting in economics and finance: A modern review. *International Journal of Forecasting*, jul 2023. doi: 10.1016/j.ijforecast.2023.05.002.

[10] Robert C. Merton. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1–2):125–144, January 1976. ISSN 0304-405X. doi: 10.1016/0304-405x(76)90022-2.

[11] S. A. Sisson, Y. Fan, and M. A. Beaumont, editors. *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC, sep 2018. doi: 10.1201/9781315117195.

[12] Seungho Yang and Jaewook Lee. Multi-basin particle swarm intelligence method for optimal calibration of parametric lévy models. *Expert Systems with Applications*, 39(1):482–493, January 2012. ISSN 0957-4174. doi: 10.1016/j.eswa.2011.07.039.

[13] Seungho Yang, Younhee Lee, Gabjin Oh, and Jaewook Lee. Calibrating parametric exponential lévy models to option market data by incorporating statistical moments priors. *Expert Systems with Applications*, 38(5):4816–4823, May 2011. ISSN 0957-4174. doi: 10.1016/j.eswa.2010.09.164.

Tarification de contrats d'assurance via des données de marchés: Application à l'assurance santé animale

Pierre-Olivier Goffard [1] & Gareth Peters [2] & Pierrick Piette [3]

[1] *Université de Strasbourg, Institut de Recherche Mathématique Avancée, Strasbourg, France, Email: goffard@unistra.fr.*

[2] *University of California Santa Barbara, Department of Statistics and Applied Probability, Santa Barbara CA 93106-3110, USA. Email: garethpeters@ucsb.edu*

[3] *Univ Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances (ISFA), Laboratoire SAF EA2429, F-69366, Lyon, France. Email: pierrick.piette@gmail.com*

Résumé. Nous présentons une méthode pour tarifer les polices d'assurance en utilisant des données de marché. L'approche est conçue pour des situations dans lesquelles la compagnie d'assurance cherche à entrer sur un nouveau marché, ici : l'assurance pour animaux de compagnie, en l'absence de données historiques. La méthodologie repose sur un processus itératif en deux étapes. Premièrement, une valeur de paramètre est proposée pour caractériser le risque sous-jacent. Deuxièmement, la prime pure résultante est reliée à la prime commerciale observée à l'aide d'un modèle de régression isotonique.

Des tests sont réalisés sur des données synthétiques, suivis d'une application sur un jeu de données réel composées de tarifs d'assurance pour animaux de compagnie récoltés en ligne. Afin de faciliter la mise en œuvre pratique, nous avons développé un package R appelé `IsoPriceR`. En répondant au défi de la tarification des polices d'assurance en l'absence de données historiques, cette méthode contribue à améliorer les stratégies de tarification sur les marchés émergents.

Mots-clés. Tarification des assurances, Inférence bayésienne, Calcul bayésien approximatif, Régression isotonique.

Abstract. We introduce a method for pricing insurance policies using market data. The approach is designed for scenarios in which the insurance company seeks to enter a new market, in our case: pet insurance, lacking historical data. The methodology involves an iterative two-step process. First, a suitable parameter is proposed to characterize the underlying risk. Second, the resulting pure premium is linked to the observed commercial premium using an isotonic regression model. To validate the method, comprehensive testing is conducted on synthetic data, followed by its application to a dataset of actual pet insurance rates. To facilitate practical implementation, we have developed an R package called `IsoPriceR`. By addressing the challenge of pricing insurance policies in the absence of historical data, this method contributes to enhancing pricing strategies in emerging markets.

Keywords. Insurance Pricing, Bayesian Inference, Approximate Bayesian Computation, Isotonic Regression.

1 Introduction

La tarification des contrats d'assurances repose sur des méthodes prédictives pour que les primes reflètent le coût moyen des sinistres. Les assureurs s'appuient généralement sur des données historiques afin d'entraîner des modèles statistiques pour estimer la prime pure. Un problème se pose lorsqu'une compagnie d'assurance entre sur un nouveau marché et ne possède pas de données historiques sur les risques qu'elle souhaite couvrir.

Nous proposons une analyse des données de marché, constituées des tarifs proposés par les concurrents pour des polices d'assurance similaires. Notre approche exploite ces données de marché pour fournir des informations sur le risque sous-jacent, conduisant au calcul des primes d'assurance.

Le processus de collecte des données implique d'obtenir des devis d'assurance. Pour recueillir ces devis, nous avons visité les sites web des compagnies d'assurance et répondu à des questionnaires concernant le risque assuré. Ce travail se focalise sur l'assurance santé des animaux de compagnie (assurance chien/chat). La méthode peut néanmoins s'adapter à d'autres produits d'assurance. L'assurance santé animale couvre les frais de santé des animaux de compagnie. Le propriétaire d'un animal doit spécifier l'espèce, la race, l'âge et le sexe de l'animal. Ces caractéristiques sont autant de facteurs tarifaires permettant la classification des risques. Une compagnie d'assurance en entrant sur un nouveau marché identifierait naturellement ces facteurs de risque lors de la collecte de données sur les primes proposées par les concurrents.

Dans une classe de risque spécifique, le risque est représenté par une variable aléatoire positive, notée X, qui représente le montant total des sinistres sur la période couverte par la police d'assurance. Les compagnies d'assurance atténuent ce risque en offrant une couverture pour une portion $g(X) < X$, en échange d'une prime. Le processus implique de calculer la prime pure, définie comme $p = \mathbb{E}[g(X)]$. Les clients se voient ensuite proposer une prime commerciale liée à la prime pure par l'équation $\widetilde{p} = f(p) > p$, où f représente la fonction de chargement. Notre problème est formulé comme suit : étant donné une collection de devis d'assurance $\mathcal{D} = \{\widetilde{p}_1, \ldots, \widetilde{p}_n\}$ associés à un risque spécifique et à des fonctions de chargement et de couverture variables données par

$$\widetilde{p}_i = f_i\{\mathbb{E}[g_i(X)]\},\ i = 1, \ldots, n,$$

peut-on identifier les paramètres de la distribution du risque sous-jacent X ainsi qu'à approximer les fonctions de chargement afin de proposer nos propres devis d'assurance?

Notre approche suppose que la distribution du risque X est paramétrée par $\theta \in \mathbb{R}^d$. Nous adoptons une procédure en trois étapes. Nous échantillonnons θ à partir d'une distribution a priori $\pi(\theta)$ qui délimite notre espace de paramètres. Nous calculons les primes pures $p_i^\theta = \mathbb{E}_\theta[g_i(X)]$, basées sur les couvertures g_i connues fournies par la police d'assurance et la valeur de paramètre θ échantillonnée. Un verrou est le manque d'expressions exploitables pour les primes pures, car nous supposons que X suit une distribution composée (c'est-à-dire une somme aléatoire). L'évaluation des primes pures passe par des simulations, ce qui entraîne une erreur d'approximation liée au nombre de réplications utilisées. Les fonctions de chargement $f_i : \mathbb{R}_+ \to \mathbb{R}_+$ sont approximées à l'aide d'un modèle de régression isotonique, choisi pour

sa capacité à maintenir la relation monotone entre les primes pures et commerciales. La procédure peut être résumée comme suit :

1. Échantillonner une valeur de paramètre θ
2. Calculer les primes pures $p_i^{\theta^*}$ pour chaque police d'assurance $i = 1, \ldots, n$
3. Ajuster une fonction isotonique f pour apprendre la relation entre les primes commerciales $\widetilde{p}_i$ et les primes pures p_i^{θ}
4. Construire les données de marché 'synthétiques' $\mathcal{D}^{\theta}$ en appliquant le chargement estimé f à la prime pure $f(p_i^{\theta})$ pour $i = 1, \ldots, n$
5. Si les données de marché observées et synthétiques sont suffisamment proches, alors nous stockons la valeur de paramètre θ et la fonction de chargement associée f.

Après avoir itéré ces étapes, nous obtenons une suite de paires paramètre-fonction de chargement : $(\theta_1, f_1), (\theta_2, f_2), \ldots$ nous permettant de tarifer nos propres polices d'assurance. Le problème que nous abordons est un problème inverse et notre solution s'inspire des méthodologies d'inférence indirecte introduites par Gourieroux et al. [6]. L'algorithme proposé pour explorer l'espace des paramètres ressemble aux algorithmes de Calcul Bayésien Approximatif (ABC) décrits dans l'ouvrage de Sisson et al. [9]. Les valeurs de paramètres échantillonnées $\theta_1, \theta_2, \ldots$ par l'algorithme fournissent une distribution postérieure approximative $\pi(\theta|\mathcal{D})$. Cette distribution postérieure prend en compte l'incertitude autour de la valeur estimée du paramètre en raison de l'utilisation de simulations pour calculer la prime pure. Les algorithmes ABC ont permis de résoudre divers problèmes en science actuarielle et en gestion des risques. Nous renvoyons les lecteurs aux travaux de Peters et al. [8], Dean et al. [3], Peters and Sisson [7] et Goffard and Laub [4] pour plus de détails. La régression isotonique est une méthode statistique bien établie, voir par exemple Barlow et al. [1], qui joue un rôle central dans notre approche.

2 Calcul de la prime d'assurance

Un individu cherche à se couvrir contre un risque X modélisé par une variable aléatoire positive, sur une période donnée, définie par

$$X = \sum_{k=1}^{N} U_k, \tag{1}$$

où N est une variable aléatoire de comptage et les U_k sont des variables aléatoires positives indépendantes et identiquement distribuées (IID), indépendantes de N. La variable aléatoire N représente le nombre d'occurrences d'un événement sur une période donnée (annuellement), chacune de ces occurrences étant associé à une indemnisation U_k. Nous supposons ici que X représente un risque appartenant à une classe de risque spécifique déterminée par des facteurs de risque.

2.1 Calcul de la prime pure

Une compagnie d'assurance propose de prendre en charge une partie de ce risque $g(X) \leq X$ en échange d'une prime qui devrait compenser le coût moyen des sinistres, donné par $p = \mathbb{E}\left[g(X)\right]$, appelée la prime pure. Dans ce travail, nous considérons une fonction g définie comme

$$g(x) = \min(\max(r \cdot x - d,\, 0),\, l),$$

où $r \in (0, 1]$ est le taux de couverture, $d > 0$ est la franchise et $l > 0$ est la limite.

2.2 De la prime pure à la prime commerciale

Soit $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ une fonction croissante telle que

$$\widetilde{p} = f(p) \geq p.$$

La fonction f est appelée fonction de chargement, un exemple simple est une fonction linéaire en la prime pure, avec

$$f(x) = (1 + \eta)x,$$

où $\eta > 0$. Les fonctions de chargement utilisées par les compagnies d'assurance nous sont inconnues et varient d'une compagnie à l'autre.

3 Tarification via les données de marché

Les données disponibles sont une collection de devis d'assurance $\widetilde{p}_{1:n} = \{\widetilde{p}_1, \ldots, \widetilde{p}_n\}$ associés à des couvertures $g_1, \ldots, g_n$ pour un risque donné X appartenant à une classe de risques spécifique. Nous supposons que ces deis d'assurance $\mathcal{D}$ proviennent du modèle décrit dans Hypothèse 1.

Hypothèse 1. *Les primes commerciales $\widetilde{p}_{1:n} = \{\widetilde{p}_1, \ldots, \widetilde{p}_n\}$ sont donnés par*

$$\widetilde{p}_i = f_i(p_i) = f_i\left\{\mathbb{E}_{\theta_0}\left[g_i(X)\right]\right\},\ i = 1, \ldots, n,$$

où le risque X est une variable aléatoire définie dans (1), paramétrée par un paramètre inconnu $\theta_0 \in \Theta \subset \mathbb{R}^d$. Les fonctions de chargement $f_1, \ldots, f_n$ sont inconnues. Les fonctions de couverture d'assurance $g_1, \ldots, g_n$ sont connues et de la forme

$$g_i(x) = \min(\max(r_i \cdot x - d_i,\, 0),\, l_i),\ \textit{pour } i = 1, \ldots, n.$$

Nous souhaitons identifier le paramètre θ qui explique le mieux nos données en nous basant sur une fonction de chargement f qui tente de moyenner les fonctions de chargement $f_1, \ldots, f_n$ utilisées par les concurrents. Cela nous conduit à formuler le problème d'optimisation suivant:

Problème 1. *Nous souhaitons trouver $\theta \in \Theta \subset \mathbb{R}^d$ et une fonction $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ afin de minimiser $d\left[\widetilde{p}_{1:n}, f(p^\theta_{i:n})\right]$, où la fonction f est appliquée élément par élément sur $p^\theta_{i:n}$, et $d(\cdot, \cdot)$ désigne une fonction de distance dans l'espace des observations, sous les contraintes suivantes:*

$$\widetilde{p}_i \geq p_i^\theta \tag{2}$$

et

$$f(p_i^\theta) \geq p_i^\theta, \tag{3}$$

pour $i = 1, \ldots, n$.

Notre solution consiste à proposer des valeurs de paramètres $\theta \in \Theta$ pour évaluer $p_{1:n}^\theta$. La fonction f est obtenue en calibrant un modèle de régression isotonique. L'objectif est alors de trouver θ qui minimise

$$d\left[\widetilde{p}_{1:n}, f(p_{1:n}^\theta)\right] = \text{RMSE}\left[\widetilde{p}_{1:n}, f(p_{1:n}^\theta)\right] + \text{Reg}_{\text{low}}\left(\widetilde{p}_{1:n}, p_{1:n}^\theta\right) + \text{Reg}_{\text{high}}\left(\widetilde{p}_{1:n}, p_{1:n}^\theta\right).$$

où

$$\text{RMSE}\left[\widetilde{p}_{1:n}, f(p_{i:n}^\theta)\right] = \sqrt{\sum_{i=1}^{n} w_i^{\text{RMSE}}[\widetilde{p}_i - f(p_i^\theta)]^2},$$

$$\text{Reg}_{\text{low}}\left(\widetilde{p}_{1:n}, p_{1:n}^\theta\right) = \sqrt{\sum_{i=1}^{n} w_i^{\text{RMSE}}\left(\widetilde{p}_i - p_i^\theta \cdot \text{LR}_{\text{low}}^{-1}\right)_+^2},$$

et

$$\text{Reg}_{\text{high}}\left(\widetilde{p}_{1:n}, p_{1:n}^\theta\right) = \sqrt{\sum_{i=1}^{n} w_i^{\text{RMSE}}\left(p_i^\theta \cdot \text{LR}_{\text{high}}^{-1} - \widetilde{p}_i\right)_+^2},$$

où $(x)_+ = \max(x, 0)$ est la partie positive de x. Les termes de régularisations $\text{Reg}_{\text{low}}\left(\widetilde{p}_{1:n}, p_{1:n}^\theta\right)$ et $\text{Reg}_{\text{high}}\left(\widetilde{p}_{1:n}, p_{1:n}^\theta\right)$ sont ajoutés au RMSE pour palier au problème d'identifiabilité du modèle (plusieurs jeux de paramètres résolvent le problème). Le ratio $p/\widetilde{p}$ correspond à ce que les praticiens appellent le rapport sinistre sur prime (S/P) attendu[1]. Nous imposons via la régularisation que les S/P $\text{LR}_i = p_i/\widetilde{p}_i$, pour $i = 1, \ldots, n$ se situent dans l'intervalle $[\text{LR}_{\text{low}}, \text{LR}_{\text{high}}]$. En supposant que $\text{LR}_{\text{high}} < 1$, nous garantissons en passant les contraintes (2) et (3). Le problème d'identifiabilité n'est pas entièrement résolu par l'ajout des termes de régularisations. Nous utilisons une extension de l'algorithme décrit dans Section 1 appelé Population Monte Carlo, voir Beaumont et al. [2]. La solution est un nuage de particules pondérées qui donne une idée de l'ensemble des paramètres admissibles. L'incertitude lié à l'évaluation des primes pures par simulation est naturellement prise en compte par l'interprétation bayésienne de notre algorithme qui correspond à un algorithme ABC.

4 Application sur un jeu de données réelles

Nous disposons de 90 devis provenant de 5 assureurs opérant sur le marché de l'assurance santé animale, pour une classe de risque spécifique associée à une femelle Berger Australien de 4 ans. Nous supposons que la fréquence suit une loi de Poisson $\mathsf{Poisson}(\lambda)$ et que la sévérité suit une loi Lognormale $\mathsf{LogNorm}(\mu, \sigma = 1)$. Les paramètres suivent les lois a priori suivantes:

$$\lambda \sim \mathsf{Unif}([0, 10]), \text{ et } \mu \sim \mathsf{Unif}([-10, 10]). \tag{4}$$

Les bornes pour les rapports S/P sont fixées à $\text{LR}_{\text{low}} = 40\%$ et $\text{LR}_{\text{high}} = 70\%$[2]. La distribution a posteriori des paramètres est donnée par la Figure 1.

[1] *Loss Ratio* en anglais noté LR

[2] fixé par avis d'expert

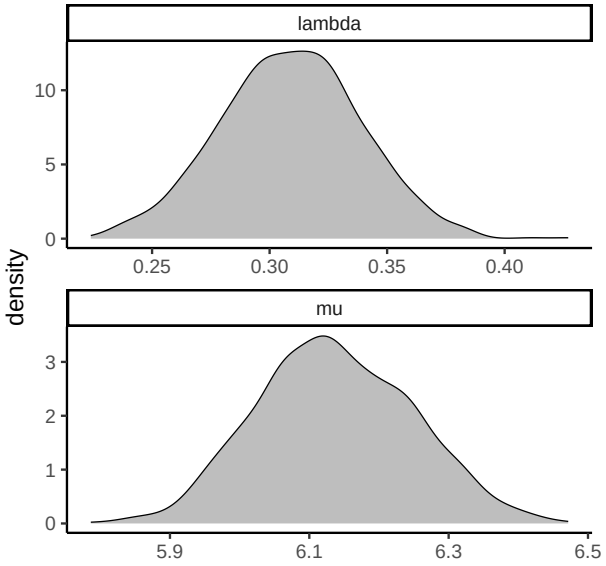


Figure 1: Distribution a posteriori des paramètres du modèle $\mathsf{Poisson}(\lambda) - \mathsf{LogNorm}(\mu, \sigma = 1)$.

L'algorithme met à jour la distribution a priori de manière informative. Table 1 rapporte les estimations des paramètres en utilisant la moyenne a posteriori (MAP) et le mode a posteriori (MODE). La Figure 2 montre un graphique illustrant la relation entre la prime commerciale

Model		MAP	MODE
$\mathsf{Poisson}(\lambda) - \mathsf{LogNorm}(\mu, \sigma = 1)$	λ	0.31	0.30
	μ	6.14	6.19

Table 1: MAP and MODE estimator for the parameters of the loss models.

et la prime pure pour le modèle $\mathsf{Poisson}(\lambda) - \mathsf{LogNorm}(\mu, \sigma = 1)$. Les différentes compagnies d'assurance sont indiquées par des couleurs distinctes, offrant une représentation visuelle des tarifs respectifs de chaque compagnie. La précision de l'ajustement du modèle de pertes nous permet de condenser les informations tridimensionnelles sur le taux de couverture, la franchise et la limite en une seule métrique : la prime pure. Ensuite, la régression isotonique révèle la relation entre les primes commerciales et pures. Les acteurs du marché de l'assurance pour animaux de compagnie est mis en exergue par les points de couleur, offrant un aperçu des stratégies tarifaires adoptées par les participants du secteur. Un preprint est disponible sur HAL, voir Goffard et al. [5].

Remerciements

Le travail de Pierre-O est réalisé dans le cadre de la Chaire de Recherche DIALOG, sous l'égide de la Fondation du Risque, une initiative de CNP Assurances. Il a aussi reçu le soutien du projet ANR DREAMES.

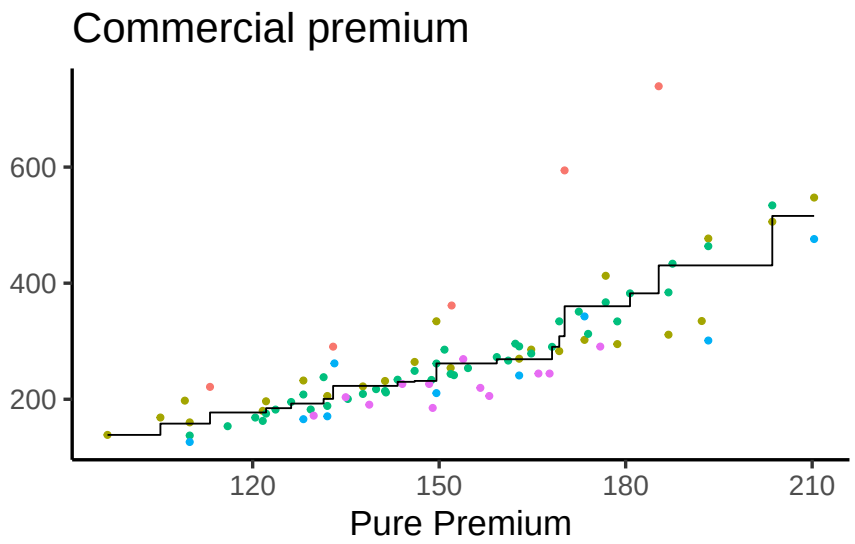


Figure 2: Prime commerciale en fonction de la prime pure pour le modèle $\mathsf{Poisson}(\lambda) - \mathsf{LogNorm}(\mu, \sigma = 1)$ en fonction de la compagnie d'assurance.

References

[1] Richard E Barlow, HD Brunk, Daniel J Bartholomew, and James M Bremner. *Statistical inference under order restrictions.(the theory and application of isotonic regression)*. 1972.

[2] Mark A Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009.

[3] Thomas A Dean, Sumeetpal S Singh, Ajay Jasra, and Gareth W Peters. Parameter estimation for hidden markov models with intractable likelihoods. *Scandinavian Journal of Statistics*, 41(4): 970–987, 2014.

[4] Pierre-Olivier Goffard and Patrick J. Laub. Approximate bayesian computations to fit and compare insurance loss models. *Insurance: Mathematics and Economics*, 100:350–371, sep 2021. doi: 10.1016/j.insmatheco.2021.06.002.

[5] Pierre-Olivier Goffard, Pierrick Piette, and Gareth W. Peters. Market-based insurance ratemaking. working paper or preprint, 2023. URL https://hal.science/hal-04297811.

[6] C. Gourieroux, A. Monfort, and E. Renault. Indirect inference. *Journal of Applied Econometrics*, 8(S1):S85–S118, dec 1993. doi: 10.1002/jae.3950080507.

[7] Gareth Peters and Scott Sisson. Bayesian inference, monte carlo sampling and operational risk. *Peters GW and Sisson SA (2006)"Bayesian Inference, Monte Carlo Sampling and Operational Risk". Journal of Operational Risk*, 1(3), 2006.

[8] Gareth W Peters, Mario V Wüthrich, and Pavel V Shevchenko. Chain ladder method: Bayesian bootstrap versus classical bootstrap. *Insurance: Mathematics and Economics*, 47(1):36–51, 2010.

[9] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC, 2018.

Combination of traditional and parametric insurance: calibration method based on the optimization of a criterion adapted to heavy tail losses.

Olivier Lopez [1] & Daniel Nkameni [1,2]

[1] *CREST Laboratory, France; E-mail: olivier.lopez@ensae.fr*
[2] *Detralytics, France; E-mail: daniel.nkameni@ensae.fr*

Résumé. Dans cet article, nous abordons la question de fournir une protection d'assurance contre les pertes à queue lourde, pour lesquelles l'espérance de la perte peut même ne pas être finie. Le produit que nous étudions repose sur une combinaison d'une assurance traditionnelle jusqu'à une certaine limite et d'une couverture paramétrique (ou basée sur un indice) pour les pertes plus importantes. Cette deuxième partie de la couverture est calculée à partir de covariables disponibles immédiatement après le sinistre, ce qui permet de réduire les coûts de gestion des sinistres grâce à une indemnisation instantanée. Pour optimiser la conception de cette seconde partie du produit, nous utilisons un critère visant à refléter les préférences de l'assuré et adapté aux pertes extrêmes (c'est-à-dire une distribution des pertes de type Pareto). Nous étayons la procédure de calibration par des résultats théoriques montrant son taux de convergence, ainsi que par des résultats empiriques issus d'une étude de simulation et d'une analyse de données réelles sur les tornades aux États-Unis.

Mots-clés. Assurance paramétrique, distributions à queue lourde, catastrophes naturelles, tarification par utilité espérée.

Abstract. In this paper, we consider the question of providing insurance protection against heavy tail losses, where the expectation of the loss may not even be finite. The product we study is based on a combination of traditional insurance up to some limit, and a parametric (or index-based) cover for larger losses. This second part of the cover is computed from covariates available just after the claim, allowing to reduce the claim management costs via an instant compensation. To optimize the design of this second part of the product, we use a criterion which aims to reflect the preferences of the policyholder, and is adapted to extreme losses (that is distribution of the losses that are of Pareto type). We support the calibration procedure by theoretical results that show its convergence rate, and empirical results from a simulation study and a real data analysis on tornados in the US.

Keywords. Parametric insurance, heavy-tail distributions, natural disasters, expected utility pricing.

References

M. J. Abdi, N. Raffar, Z. Zulkafli, K. Nurulhuda, B. M. Rehan, F. M. Muharam, N. A. Khosim, and F. Tangang. Index-based insurance and hydroclimatic risk management in agriculture: A systematic review of index selection and yield-index modelling methods. *International Journal of Disaster Risk Reduction*, 67:102653, 2022.

S. E. Bodily and D. M. Coleman. A portfolio of weather risk transfer contracts efficiently reduces risk. *Climate Risk Management*, page 100332, 2021.

M. Broberg. Parametric loss and damage insurance schemes as a means to enhance climate change resilience in developing countries. *Climate Policy*, 20(6):693–703, 2020.

M. Carter, A. de Janvry, E. Sadoulet, and A. Sarris. Index insurance for developing country agriculture: a reassessment. *Annual Review of Resource Economics*, 9(1):421–438, 2017.

Z. Chen, Y. Lu, J. Zhang, and W. Zhu. Managing weather risk with a neural network-based index insurance. *Management Science*, 70(7):4306–4327, 2024.

Y.-S. Chuang and H.-L. Chang. Issues of using parametric insurance (pi) to taiwan energy-saving performance contract (espc): A snapshot. *Energy Efficiency*, 18(1):1–13, 2025.

K. Y. Clement, W. W. Botzen, R. Brouwer, and J. C. Aerts. A global review of the impact of basis risk on the functioning of and demand for index insurance. *International Journal of Disaster Risk Reduction*, 28:845–853, 2018.

S. Conradt, R. Finger, and R. Bokusheva. Tailored to the extremes: Quantile regression for index-based insurance contract design. *Agricultural economics*, 46(4):537–547, 2015.

E. Dal Moro. Towards an economic cyber loss index for parametric cover based on it security indicator: a preliminary analysis. *Risks*, 8(2):45, 2020.

E. Eze, A. Girma, A. Zenebe, J. M. Kourouma, and G. Zenebe. Exploring the possibilities of remote yield estimation using crop water requirements for area yield index insurance in a data-scarce dryland. *Journal of Arid Environments*, 183:104261, 2020.

R. Figueiredo, M. L. Martina, D. B. Stephenson, and B. D. Youngman. A probabilistic paradigm for the parametric insurance of natural hazards. *Risk Analysis*, 38(11):2400–2414, 2018.

W. K. Härdle and B. L. Cabrera. Calibrating cat bonds for mexican earthquakes. *Journal of Risk and Insurance*, 77(3):625–650, 2010.

J. B. Horton. Parametric insurance as an alternative to liability for compensating climate harms. *Carbon & Climate Law Review*, 12(4):285–296, 2018.

L. Johnson. Paying ex gratia: Parametric insurance after calculative devices fail. *Geoforum*, 125:120–131, 2021.

X. Lin and W. J. Kwon. Application of parametric insurance in principle-compliant and innovative ways. *Risk Management and Insurance Review*, 23(2):121–150, 2020.

O. Lopez and M. Thomas. Parametric insurance for extreme risks: the challenge of properly covering severe claims. *arXiv preprint arXiv:2301.07776*, 2023.

A. Louaas and P. Picard. Optimal insurance coverage of low-probability catastrophic risks. *The Geneva Risk and Insurance Review*, 46:61–88, 2021.

L. Mastroeni, A. Mazzoccoli, and M. Naldi. Pricing cat bonds for cloud service failures. *Journal of Risk and Financial Management*, 15(10):463, 2022.

E. Michel-Kerjan and F. Morlaye. Extreme events, global warming, and insurance-linked securities: How to trigger the "tipping point". *The Geneva Papers on Risk and Insurance-Issues and Practice*, 33:153–176, 2008.

OCDE. *Enhancing the Role of Insurance in Cyber Risk Management*. 2017. doi: https://doi.org/https://doi.org/10.1787/9789264282148-en. URL `https://www.oecd-ilibrary.org/content/publication/9789264282148-en`.

O. Prokopchuk, I. Prokopchuk, G. Mentel, and Y. Bilan. Parametric insurance as innovative development factor of the agricultural sector of economy. *AGRIS on-line Papers in Economics and Informatics*, 10(665-2021-565):69–86, 2020.

K. S. Tan and J. Zhang. Flexible weather index insurance design with penalized splines. *North American Actuarial Journal*, 28(1):1–26, 2024.

J. M. Van Nostrand and J. G. Nevius. Parametric insurance: using objective measures to address the impacts of natural disasters and climate change. *Environmental Claims Journal*, 23(3-4):227–237, 2011.

Graphes et réseaux

Estimation consistante du nombre de clusters dans les modèles à blocs stochastiques multi-couches et dynamiques

Lucie ARTS [1]

[1] *arts@lpsm.paris, Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France*

Résumé. Nous introduisons l'estimateur Krichevsky-Trofimov (KT) pénalisé pour estimer le nombre de clusters lors de l'observation de plusieurs graphes dans des modèles à blocs stochastiques, tant multi-couches que dynamiques. Nous établissons la consistance de l'estimateur KT pénalisé, montrant qu'il converge vers le nombre correct de clusters dans les deux types de modèles lorsque le nombre de nœuds dans les graphes augmente. Notre estimateur ne nécessite pas de borne supérieure connue sur ce nombre pour être consistant et les résultats de consistance sont valables à la fois dans des régimes denses et parcimonieux, ce qui rend l'estimateur KT pénalisé robuste dans diverses configurations de réseaux.

Mots-clés. Graphes, Estimation de l'ordre, Classification

Abstract. We present the penalized Krichevsky-Trofimov (KT) estimator for estimating the number of clusters when observing multiple graphs within both multi-layer and dynamic Stochastic Block Models. We demonstrate the consistency of this estimator, showing that it converges to the correct number of clusters in both types of models as the number of nodes in the graphs increases. Our estimator does not require a known upper bound on this number for consistency and these consistency results hold in both dense and sparse regimes, making the penalized KT estimator robust across different network configurations.

Keywords. Graphs, Order estimation, Clustering

1 Introduction

Dans ce travail, nous nous intéressons au clustering dans les réseaux, une approche visant à classifier les nœuds en groupes. Une question centrale est l'estimation de l'ordre du modèle, c'est-à-dire la détermination du nombre de clusters, une étape essentielle pour interpréter correctement la structure d'un réseau.

Les modèles à blocs stochastiques (SBM), introduits par Holland, Laskey et Leinhardt (1983), sont couramment utilisés pour modéliser les réseaux. Le choix du nombre de groupes dans ces modèles s'appuie souvent sur des critères comme l'Integrated Complete Likelihood (ICL), proposé par Biernacki, Celeux et Govaert (2000). Cependant, un défi majeur est de démontrer la consistance de ces méthodes sans supposer une borne supérieure a priori sur le nombre de clusters.

Récemment, une avancée importante a été réalisée par Cerqueira et Leonardi (2020) grâce à l'utilisation de l'estimateur de Krichevsky-Trofimov (KT), qui permet une estimation robuste sans nécessiter de borne supérieure sur le nombre de clusters. Elles ont démontré la consistence de l'estimateur de Krichevsky-Trofimov pénalisé pour estimer le nombre de clusters dans les SBM. Ce résultat est valable aussi bien dans les régimes denses que dans les régimes parcimonieux. Cependant, ces travaux se limitent aux SBMs uniques et ne couvrent pas les collections de graphes.

Ce travail élargit cette estimation à des collections de graphes sur un même ensemble de nœuds, dans des modèles multi-couches (MLSBM) et dynamiques (DynSBM). Le MLSBM définit par Han, XU et Airoldi (2015) analyse des graphes ayant la même structure latente observée simultanément. Le DynSBM définit par Matias et Miele (2017) observe l'évolution des groupes de nœuds au fil du temps et combine un SBM statique et des chaînes de Markov indépendantes pour modéliser l'évolution temporelle des groupes. La consistance de l'estimateur KT pénalisé est établie dans les deux types de modèles lorsque le nombre de nœuds dans les réseaux augmente. Cet estimateur ne nécessite pas de borne supérieure connue sur le nombre de clusters pour être consistant et les résultats de consistance sont valables à la fois dans des régimes denses et dans des régimes parcimonieux, ce qui rend l'estimateur KT pénalisé robuste dans diverses configurations de réseaux.

2 Définitions et notations

2.1 Les modèles de collections de graphes

Nous commençons par les définitions des modèles multi-couches et dynamiques, deux modèles de collections de graphes. Il sont définis pour des graphes simples non orientés pour simplifier les notations, mais une généralisation aux autres types de graphes est aisée.

Considérons une collection de graphes simples non-orientés formée de n noeuds et indexée par $t = 1, \dots, T$ (couche ou temps). On observe des matrices d'adjacence $\mathbf{A}_{n\times n}^{1:T} = \left(A_{n\times n}^1, \dots, A_{n\times n}^T\right)$ où chaque matrice $A_{n\times n}^t$ est symétrique. On suppose qu'à chaque t les noeuds sont répartis en k groupes latents donnés par $\mathbf{Z}_n^{1:T} = (Z_i^t)_{1\le t\le T, 1\le i\le n}$ où $Z_i^t \in \{1, \dots, k\}$ est le groupe du sommet i au moment t. On suppose aussi que conditionnellement aux $\{Z_i^t\}_{1\le t\le T, 1\le i\le n}$ les graphes sont indépendants. Enfin nous supposons que

$$A_{ij}^t \mid \{Z_i^t = a, Z_j^t = b\} \sim \mathcal{B}(P_{ab}^t).$$

On introduit les compteurs suivants pour un graphe observé $a_{n\times n}$ et pour une configuration latente $\mathbf{z}_n$

$$n_a(\mathbf{z}_n) = \sum_{i=1}^{n} \mathbb{1}\{z_i = a\}, \quad 1 \le a \le k,$$

$$n_{ab}(\mathbf{z}_n) = \begin{cases} n_a(\mathbf{z}_n) n_b(\mathbf{z}_n), & 1 \le a < b \le k, \\ \frac{1}{2} n_a(\mathbf{z}_n)(n_a(\mathbf{z}_n) - 1), & 1 \le a = b \le k, \end{cases}$$

$$o_{ab}(\mathbf{z}_n, a_{n\times n}) = \begin{cases} \sum_{1\le i,j\le n} \mathbb{1}\{z_i = a, z_j = b\}\, a_{ij}, & a < b, \\ \sum_{1\le i<j\le n} \mathbb{1}\{z_i = a, z_j = b\}\, a_{ij}, & a = b, \end{cases}$$

et

$$c_{ab}(\mathbf{z}_n^{1:T}) = \sum_{t=1}^{T-1}\sum_{i=1}^{n} \mathbb{1}\{z_i^t = a, z_i^{t+1} = b\}.$$

On a donc $n_a(\mathbf{z}_n)$ (respectivement $n_{ab}(\mathbf{z}_n)$) qui est le nombre de nœuds (resp. de paires de nœuds) dans le cluster a (resp. dans les clusters (a,b)), tandis que $o_{ab}(\mathbf{z}_n, a_{n\times n})$ est le nombre d'arêtes entre les clusters a et b, et $c_{ab}(\mathbf{Z}_n^{1:T})$ compte le nombre de transitions de nœuds du cluster a au cluster b. On va maintenant donner les spécificités de chaque modèle.

Modèle à blocs stochastiques multi-couches (MLSBM). Les variables aléatoires Z_i^t sont supposées indépendantes de t, et chaque nœud i est associé à une variable latente $Z_i \in \{1,\dots,k\}$ avec une probabilité π_i. La distribution du couple $\left(\mathbf{Z}_n, \mathbf{A}_{n\times n}^{1:T}\right)$ est donnée par

$$\left(\mathbf{z}_n, \mathbf{a}_{n\times n}^{1:T}\right) \mapsto \mathbb{P}_{\pi,\mathbf{P}}(\mathbf{z}_n, \mathbf{a}_{n\times n}^{1:T}) = \prod_{1\le a\le k} \pi_a^{n_a} \prod_{t=1}^{T} \prod_{1\le a\le b\le k} \left(P_{ab}^t\right)^{o_{ab}^t} \left(1 - P_{ab}^t\right)^{n_{ab} - o_{ab}^t}. \tag{1}$$

Nous notons $\Theta_{\mathrm{ML}}^{k,T}$ l'espace des paramètres, donné par

$$\Theta_{\mathrm{ML}}^{k,T} = \{(\pi, \mathbf{P}) : \pi \in (0,1)^k, \sum_{a=1}^{k} \pi_a = 1, \forall t \in \{1,\dots,T\}, P^t \in [0,1]^{k\times k}, P^t \text{ est symétrique}\}.$$

Modèle dynamique à blocs stochastiques (DynSBM). Les variables aléatoires $\{Z_i^{1:T}\}_{1\le i\le n}$ sont supposées indépendantes et identiquement distribuées, et chaque $Z_i^{1:T} = (Z_i^t)_{1\le t\le T}$ est une chaîne de Markov irréductible, apériodique et stationnaire avec des probabilités de transition données par

$$\mathbb{P}(Z_i^{t+1} = b \mid Z_i^t = a) = \pi_{ab}, \quad \forall 1 \le a, b \le k,$$

où $\Pi = (\pi_{ab})_{1\le a,b\le k}$ est une matrice stochastique. On note $\alpha = (\alpha_1,\dots,\alpha_k)$ la distribution initiale stationnaire de la chaîne de Markov. On a alors :

$$\mathbb{P}_{\Pi}(Z_i^{1:T}) = \alpha_{Z_i^1} \prod_{t=1}^{T-1} \pi_{Z_i^t Z_i^{t+1}}, \quad \forall i \in [\![1, n]\!].$$

Matias et Miele (2017) expliquent qu'il est nécessaire d'ajouter des contraintes sur la matrice de transition Π ou sur le paramètre $\mathbf{P}$ pour obtenir l'identifiabilité du modèle. Nous choisissons ici d'imposer que pour tout $1 \le a \le k$ et $1 \le t, t' \le T$, on ait $P_{aa}^t = P_{aa}^{t'}$. Nous notons $\Theta_{\mathrm{dyn}}^{k,T}$ l'espace des paramètres pour un DynSBM, défini par :

$$\Theta_{\mathrm{dyn}}^{k,T} = \{(\Pi, \mathbf{P}) : \Pi \in [0,1]^{k\times k} \text{ est une matrice stochastique, } \forall t \in \{1,\dots,T\}, P^t \in [0,1]^{k\times k},$$
$$P^t \text{ est symétrique, et } \forall 1 \le a \le k, \text{ et } \forall 1 \le t, t' \le T, \text{ on a } P_{aa}^t = P_{aa}^{t'}\}.$$

Connaissant la configuration au temps initial $\mathbf{Z}_n^1$, la distribution de $\left(\mathbf{Z}_n^{2:T}, \mathbf{A}_{n\times n}^{1:T}\right)$ est

$$\left(\mathbf{z}_n^{2:T}, \mathbf{a}_{n\times n}^{1:T}\right) \mapsto \mathbb{P}_{\Pi,\mathbf{P}}((\mathbf{z}_n^{2:T}, \mathbf{a}_{n\times n}^{1:T}) | \mathbf{z}_n^1) = \prod_{1\le a,b\le k} \pi_{ab}^{c_{ab}} \prod_{t=1}^{T} \prod_{1\le a\le b\le k} \left(P_{ab}^t\right)^{o_{ab}^t} \left(1 - P_{ab}^t\right)^{n_{ab}^t - o_{ab}^t}. \tag{2}$$

Ordre d'un SBM. L'ordre du MLSBM (respectivement du DynSBM) est le plus petit entier k pour lequel l'égalité (1) (resp. (2)) est satisfaite pour une paire de paramètres $(\pi^0, \mathbf{P}^0) \in \Theta_{\mathrm{ML}}^{k,T}$ (resp. $(\Pi^0, \mathbf{P}^0) \in \Theta_{\mathrm{dyn}}^{k,T}$) et est noté par k_0.

2.2 L'estimateur KT pénalisé

Estimateur KT pénalisé pour le MLSBM. L'estimateur est défini par

$$\hat{k}_{\mathrm{KT}}(\mathbf{A}_{n\times n}^{1:T}) = \underset{1\le k\le n}{\arg\max} \left\{ \log \mathbf{KT}_k^T(\mathbf{A}_{n\times n}^{1:T}) - \mathrm{pen}_{\mathrm{ML}}(k,n,T) \right\}, \tag{3}$$

avec

$$\mathrm{pen}_{\mathrm{ML}}(k,n,T) = \sum_{i=1}^{k-1} \left[\frac{Ti(i+1)+i-1}{2} + 1 + \epsilon \right] \log n,$$

pour un $\epsilon > 0$ et où

$$\mathbf{KT}_k^T(\mathbf{A}_{n\times n}^{1:T}) = \mathbb{E}_{\nu_k}\left[\mathbb{P}_{\pi,\mathbf{P}}(\mathbf{A}_{n\times n}^{1:T}) \right] = \int_{\Theta_{\mathrm{ML}}^{k,T}} \mathbb{P}_{\pi,\mathbf{P}}(\mathbf{A}_{n\times n}^{1:T}) \nu_k(\pi,\mathbf{P}) d\pi d\mathbf{P},$$

est la vraisemblance intégrée par rapport à la distribution a priori définie par

$$\nu_k(\pi,\mathbf{P}) = \frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{1}{2})^k} \prod_{1\le a\le k} \pi_a^{-\frac{1}{2}} \times \prod_{t=1}^{T} \prod_{1\le a\le b\le k} \frac{1}{\Gamma(\frac{1}{2})^2} (P_{ab}^t)^{-\frac{1}{2}} (1-P_{ab}^t)^{-\frac{1}{2}}.$$

Estimateur KT pénalisé pour DynSBM. L'estimateur est défini par

$$\hat{k}_{\mathrm{KT}}(\mathbf{A}_{n\times n}^{1:T}) = \underset{1\le k\le n}{\arg\max} \left\{ \log \mathbf{KT}_k^T(\mathbf{A}_{n\times n}^{1:T}|\mathbf{z}_n^1) - \mathrm{pen}_{\mathrm{dyn}}(k,n,T) \right\}, \tag{4}$$

avec

$$\mathrm{pen}_{\mathrm{dyn}}(k,n,T) = \sum_{i=1}^{k-1} \left[\frac{i(i-1)}{2} + i + 1 + \epsilon \right] \log(nT) + \left[\frac{Ti(i-1)}{2} + 1 + \epsilon \right] \log n,$$

pour $\epsilon > 0$, et où

$$\mathbf{KT}_k^T(\mathbf{A}_{n\times n}^{1:T}|\mathbf{z}_n^1) = \mathbb{E}_{\nu_k}\left[\mathbb{P}_{\Pi,\mathbf{P}}(\mathbf{A}_{n\times n}^{1:T}|\mathbf{z}_n^1) \right] = \int_{\Theta_{\mathrm{dyn}}^{k,T}} \mathbb{P}_{\Pi,\mathbf{P}}(\mathbf{A}_{n\times n}^{1:T}|\mathbf{z}_n^1) \nu_k(\Pi,\mathbf{P}) d\Pi d\mathbf{P},$$

est la vraisemblance intégrée (conditionnelle à l'état latent au premier temps) par rapport à la distribution a priori définie par

$$\begin{aligned}\nu_k(\Pi,\mathbf{P}) = \prod_{1\le b\le k} \left[\frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{1}{2})^k} \prod_{1\le a\le k} \pi_{ab}^{-\frac{1}{2}} \right] &\times \prod_{1\le a\le k} \frac{1}{\Gamma(\frac{1}{2})^2} (P_{aa})^{-\frac{1}{2}} (1-P_{aa})^{-\frac{1}{2}} \\ &\times \prod_{t=1}^{T} \prod_{1\le a<b\le k} \frac{1}{\Gamma(\frac{1}{2})^2} (P_{ab}^t)^{-\frac{1}{2}} (1-P_{ab}^t)^{-\frac{1}{2}}.\end{aligned}$$

Nous présentons les résultats suivants dans le cas du régime sparse (pour tout $1 \leq t \leq T$, on a $P^{0,t} = \rho_n^t S^{0,t}$ avec $S^{0,t}$ une matrice indépendante de n). Cependant, les résultats peuvent être transposés au cas dense en prenant $\rho_n^t = 1$ et $S^{0,t} = P^{0,t}$, et en remplaçant la fonction $\tau(u) = u \log u - u$ par la fonction $\gamma(u) = u \log u + (1-u)\log(1-u)$.

3 Résultats

3.1 Théorème de consitance

Théorème 1. *Considérons le modèle MLSBM avec T couches et k_0 clusters (resp. le modèle DynSBM avec T temps et k_0 clusters). Soit $\hat{k}_{KT}$ défini comme dans* (3) *(resp.* (4)*). Alors, sous les régimes sparse et dense, nous avons presque sûrement lorsque $n \to \infty$, avec T fixe :*

$$\hat{k}_{KT}(\mathbf{A}_{n\times n}^{1:T}) = k_0.$$

3.2 Idée de la preuve du théorème

La preuve du théorème est coupée en deux parties : la non sur-estimation et la non sous-estimation. La preuve suit exactement le même chemin que celui de Cerqueira et Leonardi (2020).

Proposition 1 (Non sur-estimation)**.** *Soit $\mathbf{A}_{n\times n}^{1:T}$ un MLSBM k_0 avec T couches et n noeuds par couche (resp. un DynSBM avec k_0 communautés et T temps). Alors, l'estimateur $\hat{k}_{kT}\left(\mathbf{A}_{n\times n}^{1:T}\right)$ défini dans* (3) *(resp.* (4)*) ne surestime pas k_0, presque sûrement lorsque $n \to \infty$.*

Preuve. Cette proposition se prouve en montrant séparément que $\hat{k}_{KT}\left(\mathbf{A}_{n\times n}^{1:T}\right) \notin (k_0, \log n]$ et que $\hat{k}_{KT}\left(\mathbf{A}_{n\times n}^{1:T}\right) \notin (\log n, n]$, ce qui se montre en réécrivant ces ensembles de façon astucieuse et en utilisant le lemme de Borel Cantelli. □

Pour la non sous-estimation, nous avons d'abord besoin d'une définition :

Definition 1. *L'estimateur du profil de vraisemblance pour l'affectation des étiquettes sous un $k-$MLSBM, avec $\mathbf{A}_{n\times n}^{1:T}$ un échantillon d'un MLSBM d'ordre k_0, est défini comme*

$$\mathbf{z}_{n,k}^{\star} = \underset{\mathbf{z}_n \in [k]^n}{\arg\max} \sup_{(\pi,\mathbf{P})\in\Theta_{ML}^{k,T}} \mathbb{P}_{\pi,\mathbf{P}}\left(\mathbf{z}_n, \mathbf{A}_{n\times n}^{1:T}\right).$$

Dans le cas d'un DynSBM il est défini comme

$$\mathbf{z}_{n,k}^{\star} = \underset{\mathbf{z}_n^{2:T} \in [k]^{nT}}{\arg\max} \sup_{(\pi,\mathbf{P})\in\Theta_{dyn}^{k,T}} \mathbb{P}_{\pi,\mathbf{P}}\left(\mathbf{z}_n^{2:T}, \mathbf{A}_{n\times n}^{1:T} | \mathbf{z}_n^1\right).$$

Proposition 2 (Non sous-estimation)**.** *Supposons que $(\mathbf{Z}_n, \mathbf{A}_{n\times n}^{1:T})$ soit un MLSBM de n noeuds, d'ordre k_0 avec T couches, de paramètres $(\pi^0, \mathbf{P}^0)$ (resp. $(\mathbf{Z}_n^{1:T}, \mathbf{A}_{n\times n}^{1:T})$ un DynSBM d'ordre k_0 de paramètres $(\Pi^0, \mathbf{P}^0)$), avec $\mathbf{P}^0 = \boldsymbol{\rho}_n \mathbf{S}^0$ où $\mathbf{S}^0$ ne dépend pas de n, et $\boldsymbol{\rho}_n = (\rho_n^1, \ldots, \rho_n^T)$ (resp. ρ_n ne dépend pas de t) avec $\rho_n^t \geq C \log n/n$ et $\rho_n^t \to 0$. Alors, $\hat{k}_{KT}\left(\mathbf{A}_{n\times n}^{1:T}\right)$ ne sous-estime pas k_0, presque sûrement lorsque $n \to \infty$.*

Preuve. Nous commençons par le cas du MLSBM, la preuve pour le DynSBM est analogue. En suivant l'approche de Cerqueira et Leonardi (2020) dans la preuve de la Proposition 6, pour montrer que $\hat{k}_{KT}\left(\mathbf{A}_{n\times n}^{1:T}\right)$ ne sous-estime pas k_0, il suffit de démontrer que pour tout $k < k_0$,

$$\liminf_{n\to\infty} \frac{1}{\min\limits_{1\leq t\leq T} \rho_n^t n^2} \log \frac{\sup_{(\pi,\mathbf{S})\in\Theta_{\mathrm{ML}}^{k_0,T}} \mathbb{P}_{\pi,\boldsymbol{\rho}_n\mathbf{S}}\left(\mathbf{A}_{n\times n}^{1:T}\right)}{\sup_{(\pi,\mathbf{S})\in\Theta_{\mathrm{ML}}^{k,T}} \mathbb{P}_{\pi,\boldsymbol{\rho}_n\mathbf{S}}\left(\mathbf{A}_{n\times n}^{1:T}\right)} > 0.$$

Nous commençons avec $k = k_0 - 1$. En utilisant les estimateurs du maximum de vraisemblance dans le modèle complet pour π_a et $S_{a,b}^t$, en utilisant que pour $\rho_n \to 0$, nous avons que $\gamma(\rho_n x) = \rho_n \tau(x) + x\rho_n \log(\rho_n) + O\left(\rho_n^2 x^2\right)$ et en divisant par $\min\limits_{1\leq t\leq T} \rho_n^t n^2$, il nous suffit de montrer que pour tout $1 \leq t \leq T$

$$\begin{aligned}\liminf_{n\to\infty} \sum_{1\leq a,b\leq k_0} \frac{n_a(\mathbf{z}_n) n_b(\mathbf{z}_n)}{n^2} \tau\left(\frac{\tilde{o}_{a,b}^t(\mathbf{z}_n, \mathbf{A}_{n\times n}^t)}{\rho_n^t n_a(\mathbf{z}_n) n_b(\mathbf{z}_n)}\right) \\ - \sum_{1\leq a,b\leq k} \frac{n_a(\mathbf{z}_{n,k}^\star) n_b(\mathbf{z}_{n,k}^\star)}{n^2} \tau\left(\frac{\tilde{o}_{a,b}^t(\mathbf{z}_{n,k}^\star, \mathbf{A}_{n\times n}^t)}{\rho_n^t n_a(\mathbf{z}_{n,k}^\star) n_b(\mathbf{z}_{n,k}^\star)}\right) > 0,\end{aligned}$$

avec $\tilde{o}_{ab} = o_{ab}$ pour $a \neq b$, et $\tilde{o}_{aa} = 2o_{aa}$. Cette inégalité se prouve par deux lemmes omis ici. Enfin pour $k < k_0 - 1$, il suffit de décomposer ce que l'on veut démontrer en une partie identique au cas précédent et une partie qui sera positive par croissance de la vraisemblance en fonction de la taille du modèle.

□

Bibliographie

Biernacki, C. et Celeux, G. et Govaert, G (2000), Assessing a mixture model for clustering with the integrated completed likelihood, *EEE Transactions on Pattern Analysis and Machine Intelligence*, 22, pp. 719-725

Cerqueira, Andressa et Leonardi, Florencia (2020), Estimation of the Number of Communities in the Stochastic Block Model, *IEEE Transactions on Information Theory*, 66, pp. 6403-6412.

Han, Qiuyi et Xu, Kevin et Airoldi, Edoardo (2015), Consistent estimation of dynamic and multi-layer block models, *Proceedings of the 32nd International Conference on Machine Learning*, 37.

Holland, Paul W. et Laskey, Kathryn Blackmond et Leinhardt, Samuel (1983), Stochastic blockmodels: First steps, *Social Networks*, 5, pp. 109-137.

Matias, Catherine et Miele, Vincent (2017), Statistical clustering of temporal networks through a dynamic stochastic block model, *Journal of the Royal Statistical Society: Series B*, 79, pp. 1119–1141

Détection de communauté pour un modèle graphique binaire en grande dimension

Julien Chevallier [1] & Guilherme Ost [2]

[1] *Université Grenoble Alpes, France, julien.chevallier1@univ-grenoble-alpes.fr*
[2] *Federal University of Rio de Janeiro, Brésil, guilhermeost@im.ufrj.br*

Résumé. Soient N agents partitionnés en deux communautés notées $\mathcal{P}_+$ pour la partie excitatrice et $\mathcal{P}_-$ pour la partie inhibitrice. Supposons qu'ils soient reliés via un graphe d'Erdös-Rényi dirigé de paramètre inconnue $p \in]0,1[$. A chaque pas de temps $t = 1, \dots, T$, on observe l'état de chaque agent: soit il envoie un signal à ses descendants (dans le graphe dirigé) soit il reste silencieux.

Nous proposons une procédure de classification basée sur cette seule observation binaire. Elle utilise la covariance du processus sous-jacent. Dans un régime asymptotique en terms de N et T, nous prouvons que cette procédure retrouve (de manière exacte) les deux communautés.

Mots-clés. Classification, reconstruction exacte, modèle graphique, chaîne de Markov

Abstract. Let N components be partitioned into two communities, denoted $\mathcal{P}_+$ for the excitatory one and $\mathcal{P}_-$ for the inhibitory one. Assume that they are connected via a directed and weighted Erdös-Rényi random graph (DWER) with unknown parameter $p \in (0,1)$. At each time unit $t = 1, \dots, T$, we observe the state of each component: either it sends some signal to its successors (in the directed graph) or remains silent otherwise.

Based on this mere observation, a classification procedure is proposed relying on the covariance of the process. It is proved to recover exactly the two communities under some asymptotic regime in terms of N and T.

Keywords. Classification, exact recovery, graphical model, Markov chain

1 Introduction

The objective is to classify N nodes into two communities (an excitatory one and an inhibitory one) based on the mere observation of $\{0,1\}$-valued sequences for each node. This classification task is proved to be consistent in the model introduced in the article of Chevallier et al. (2024) and presented below.

Our study is related with the articles of Berthet et al. (2019), in the context of Ising graphical models, and Bunea et al. (2020), in the context of G-latent models. In both papers, the observations are assumed to be i.i.d. copies of some N-dimensional random vector Z. In the Ising graphical model, the interaction strength is βN^{-1} between sites of the

same community and αN^{-1} otherwise with $\beta > \alpha$. In the G-latent model, the community structure is imposed directly on the covariance matrix of the random vector Z. In this presentation, we go beyond the assumption that the observations are i.i.d by addressing the community detection problem for dependent variables.

2 Graphical binary model

Let $X = \{X_{i,t}, t \in \mathbb{Z}, 1 \leq i \leq N\}$ be N interacting $\{0,1\}$-valued chains. Their interaction is encoded by a random matrix $\theta = (\theta_{ij})_{1 \leq i,j \leq N}$ of i.i.d entries distributed as $\mathrm{Ber}(p)$ with $0 \leq p \leq 1$. Let $X_t = (X_{1,t}, \ldots, X_{N,t})$ denote the configuration of the system X at time t.

The dynamics of the model is as follows. Conditionally on θ, the system X evolves as a stationary Markov chain on the state space $\{0,1\}^N$ in which the conditional distribution of X_t given that $X_{t-1} = x$ is that of N independent $\mathrm{Ber}(p_{\theta,i}(x))$ random variables, $i = 1, \ldots, N$, where

$$p_{\theta,i}(x) = \mu + (1-\lambda)\left(\frac{1}{N}\sum_{j \in \mathcal{P}_+} \theta_{ij} x_j + \frac{1}{N}\sum_{j \in \mathcal{P}_-} \theta_{ij}(1-x_j)\right), \; x = (x_1, \ldots, x_N). \tag{1}$$

The sets $\mathcal{P}_+$ and $\mathcal{P}_-$ form a partition of $[N] := \{1, \ldots, N\}$ and the values $0 < \lambda < 1$ and $0 \leq \mu \leq \lambda$ are fixed. Neither the sets $\mathcal{P}_+$ and $\mathcal{P}_-$ nor the values μ and λ depend on the random environment θ.

In this model, the transition probabilities $p_{i,\theta}(x)$ defined in (1) are increasing with respect to x_j for $j \in \mathcal{P}_+$ and decreasing for $j \in \mathcal{P}_-$, so that the components $j \in \mathcal{P}_+$ have an excitatory role, whereas those belonging to $\mathcal{P}_-$ have an inhibitory role. In that respect, the classification task is to recover the two communities $\mathcal{P}_+$ and $\mathcal{P}_-$ from the mere observation of $X_1, \ldots, X_T$.

3 Classification procedure

The classification procedure is based on the one unit time-lagged covariance matrix $\Sigma^{(1)}$ whose (i,j)-entries are defined as

$$\Sigma^{(1)}_{ij} = \mathrm{Cov}_\theta(X_{i,1}, X_{j,0}), \; i,j \in [N], \tag{2}$$

where $\mathrm{Cov}_\theta(X_{i,1}, X_{j,0})$ denotes the covariance between the random variables $X_{i,1}$ and $X_{j,0}$, conditionally on θ. Since a component j has either an excitatory or inhibitory influence towards everyone else, it is natural to look at the column sums of $\Sigma^{(1)}$, that is the vector σ^N defined by, for all $j \in [N]$,

$$\sigma^N_j = \sum_{i=1}^{N} \Sigma^{(1)}_{ij}.$$

The asymptotics of σ^N can be characterized as follows.

Proposition 1. *There exist two explicit values $c_1 > 0$ and c_2 such that, for all $N \geq 2$,*

$$\sigma_j^N = \begin{cases} c_1 p + c_2 + \sum_{i=1}^N \epsilon_{ij}^N & \text{if } j \in \mathcal{P}_+, \\ -c_1 p + c_2 + \sum_{i=1}^N \epsilon_{ij}^N & \text{if } j \in \mathcal{P}_-, \end{cases} \tag{3}$$

where the errors ϵ_{ij}^N satisfy

$$\mathbb{E}\left[\max_{i,j} |\epsilon_{ij}^N|\right] \leq K\sqrt{\frac{\log(N)}{N}}, \tag{4}$$

for some constant $K > 0$.

Since c_1 above is positive, the value of σ_j^N is a good proxy to classify between $j \in \mathcal{P}_+$ and $j \in \mathcal{P}_-$. Provided that the errors ε_{ij}^N are small enough (which occurs with high probability), the coordinates of σ^N are split into two well separated clusters.

Of course, one cannot use σ^N in practice, but the standard estimator of σ^N converges sufficiently fast. In turn, we provide communities estimators $\hat{\mathcal{P}}_+$ and $\hat{\mathcal{P}}_-$ such that exact recovery is eventually almost surely feasible in the following sense.

Theorem 1. *If as $N \to \infty$ and $T \to \infty$,*

$$\sqrt{\log(N)N^{-1}} + N\log(NT)/\sqrt{T} \to 0,$$

then the probability of exact recovery goes to 1:

$$\lim_{N,T\to\infty} I\!P\left(\left\{\hat{\mathcal{P}}_+^N = \mathcal{P}_+\right\} \cap \left\{\hat{\mathcal{P}}_-^N = \mathcal{P}_-\right\}\right) = 1.$$

The asymptotic regime above is illustrated in Figure 1.

Bibliographie

Berthet Q., Rigollet P. and Srisvastava P. (2019), Exact recovery in the Ising blockmodel, *The Annals of Statistics*, 47, pp. 1805 – 1834.

Bunea F., Giraud C., Luo X., Royer M. and Verzelen N. (2020), Model assisted variable clustering: Minimax-optimal recovery and algorithms, *The Annals of Statistics*, 48, pp. 111 – 137.

Chevallier J., Löcherbach E. and Ost G. (2024). Inferring the graph of dependencies density of binary graphical models in high dimension, *Arxiv preprint*.

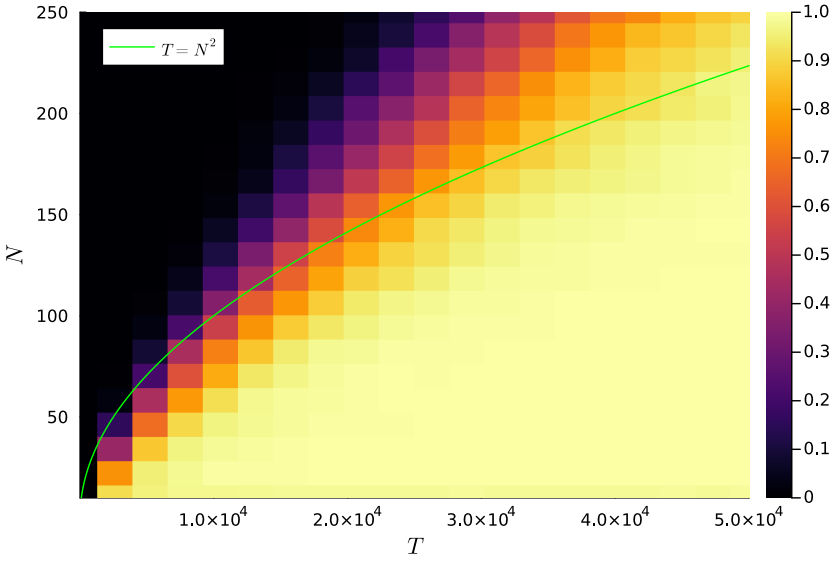


Figure 1: Estimated probability of exact recovery as a function of N and T (computed over 10^3 simulations). The green curve correspond to the equation $T = N^2$ which is the critical condition of our asymptotic regime.

Génération de modèles graphiques

Ali Fakhar[1] & Sophie Achard[2] & Irène Gannaz[3] & Kevin Polisano[4]

[1] *Univ. Grenoble Alpes, LJK, F-38000 Grenoble, France, ali.fakhar@univ-grenoble-alpes.fr*
[2] *Univ. Grenoble Alpes, CNRS, LJK, F-38000 Grenoble, France*
sophie.achard@univ-grenoble-alpes.fr
[3] *Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP, 38000 Grenoble, France*
irene.gannaz@grenoble-inp.fr
[4] *Univ. Grenoble Alpes, CNRS, LJK, F-38000 Grenoble, France*
kevin.polisano@univ-grenoble-alpes.fr

Résumé. Ce travail concerne la génération de matrices de corrélation théoriques avec des motifs de parcimonie spécifiques, associés à des structures de graphes. Nous présentons une nouvelle approche basée sur l'optimisation convexe, offrant une plus grande flexibilité par rapport aux techniques existantes. Notre méthode permet de générer des matrices de corrélation qui peuvent être utilisées pour comparer les méthodes d'inférence de graphes.

Mots-clés. Matrices de corrélation, modèles graphiques, structures de parcimonie

Abstract. This work deals with the generation of theoretical correlation matrices, with fixed sparsity patterns. We propose an optimization approach. It offers some flexibility, in particular by the addition of some constraints. The generated correlation matrices may be used for comparisons of statistical inferences in graphical models.

Keywords. Correlation matrices, graphical models, sparsity patterns

1 Introduction

Les matrices de corrélation permettent de modéliser la structure de dépendance entre des variables aléatoires. Si elles sont très présentes dans les modèles statistiques, peu de procédures de simulation sont disponibles. Simuler une matrice de corrélation n'est pas aisé en grande dimension, car il faut que la matrice créée soit définie positive. La motivation pour générer des matrices de corrélations est de permettre de simuler des jeux de données complexes afin de vérifier les performances de méthodes statistiques par Monte-Carlo. Des exemples d'application sont la détection de clusters ou l'inférence de modèles graphiques. Dans ces deux exemples, la matrice générée doit respecter certaines contraintes. Principalement, elle doit être associée à une structure de graphe sous-jacente, que les procédures essaient d'estimer.

2 Notations

Nous considérons dans la suite une dimension p fixée. Une matrice symétrique réelle $\mathbf{A} \in \mathbb{R}^{p\times p}$ est semi-définie positive (PSD) si $\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} \geq 0$ pour tout $\boldsymbol{x} \in \mathbb{R}^p$. Elle est définie positive (PD) si $\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} > 0$ pour tout $\boldsymbol{x} \in \mathbb{R}^p$ non nul.

Soit $\boldsymbol{x} \in \mathbb{R}^p$ un vecteur aléatoire de matrice de covariance $\mathbf{\Sigma}$, défini comme : $\mathbf{\Sigma} = \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^\top] = (\sigma_{i,j})_{i,j=1,\dots,p}$. La matrice de corrélation correspondante $\mathbf{C} \in \mathcal{C}$ est définie par $c_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$ pour tout $(i,j) \in \{1,\dots,p\}$, ou de manière équivalente $\mathbf{C} = (\mathrm{diag}(\mathbf{\Sigma}))^{-\frac{1}{2}}\ \mathbf{\Sigma}\ (\mathrm{diag}(\mathbf{\Sigma}))^{-\frac{1}{2}}$. L'ensemble $\mathcal{C}$ de matrices de corrélation satisfait

$$\forall \mathbf{C} \in \mathcal{C} \quad \mathrm{diag}(\mathbf{C}) = 1,\ \forall i,j \in \{1,2,\dots,p\}\ -1 \leq c_{ij} \leq 1. \tag{1}$$

D'un point de vue génératif, travailler avec des matrices de corrélation et des matrices de précision est similaire. Par conséquent, nous nous concentrons sur les matrices de corrélation. La génération d'une matrice de corrélation implique la construction d'une matrice PSD symétrique qui satisfait la condition (1) [5, Problème 7.1.].

Un graphe $\mathcal{G}$ est défini comme $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, où $\mathcal{V} = \{v_1, v_2, \dots, v_p\}$ est l'ensemble des sommets, et $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ est l'ensemble des arêtes. Une arête est une paire de sommets connectés par le graphe. La densité du graphe est alors $d = \frac{|\mathcal{E}|}{p(p-1)}$.

Nous désignons par $\mathcal{C}(\mathcal{G})$ l'ensemble des matrices de corrélation associées à un graphe donné $\mathcal{G}$. Une matrice $\mathbf{C} \in \mathcal{C}(\mathcal{G})$ est telle que

$$\mathbf{C} = (c_{ij}) \text{ PSD, satisfait (1) et } c_{ij} = 0 \text{ ssi } (i,j) \notin \mathcal{E}. \tag{2}$$

Nous recherchons à générer une matrice de corrélation dans $\mathcal{C}(\mathcal{G})$. En d'autres termes, nous imposons la parcimonie à certaines relations. Nous souhaitons générer des matrices dans $\mathcal{C}(\mathcal{G})$. Ce problème peut être considéré comme un problème de complétion de matrice [9, Chapitre 10].

3 Méthodes existantes

La génération d'une matrice de corrélation (théorique) a été proposée par exemple dans [6] avec une caractérisation de la distribution uniforme sur l'espace des matrices de corrélation. Les approches courantes incluent les *vines* et *onion* [7]. Pour une revue plus large des techniques disponibles, nous renvoyons aux études bibliographiques dans [8, 4]. La plupart des méthodes existantes ne peuvent cependant pas être étendues pour générer des matrices de corrélation avec un ensemble de parcimonie $\overline{\mathcal{E}}$ contraint. Ci-dessous, nous présentons les méthodes qui, à notre connaissance, peuvent générer de telles matrices.

3.1 Graphes chordaux et décomposition de Cholesky

La première approche repose sur la décomposition de Cholesky. Soit $\mathcal{U} \in \mathbb{R}^{p\times p}$ l'ensemble des matrices triangulaires supérieures, à éléments diagonaux positifs et à lignes normalisées

à 1. On définit $\mathcal{U}(\mathcal{G}) \subset \mathcal{U}$ comme le sous-ensemble où $u_{ij} = 0$ pour tout $(i,j) \in \overline{\mathcal{E}}$. Si le graphe est ordonné (Perfect Elimination Order, PEO), il est possible de générer le facteur de Cholesky $\mathbf{U}$ dans $\mathcal{U}(\mathcal{G})$ en imposant $u_{ij} = 0$ pour tout $(i,j) \in \overline{\mathcal{E}}$ pour obtenir une matrice $\mathbf{C} = \mathbf{U}\mathbf{U}^\top \in \mathcal{C}(\mathcal{G})$ telle que $m_{ij} = 0$ pour tout $(i,j) \notin \mathcal{E}$. Pour plus de détails, nous nous référons à [4].

Dans [8], les auteurs proposent une écriture en coordonnées polaires des éléments de la matrice de Cholesky $\mathbf{U} \in \mathcal{U}$. Ils établissent la distribution de probabilité des quantités induites, telle que la distribution des matrices obtenues soit uniforme sur l'ensemble $\mathcal{C}$. En utilisant cette paramétrisation en coordonnées polaires, il est aisé d'incorporer la contrainte $u_{ij} = 0$ pour tout $(i,j) \in \overline{\mathcal{E}}$, pour obtenir $\mathbf{U} \in \mathcal{U}(\mathcal{G})$.

Dans [4] la première contrainte est déjà prise en compte. La méthode de génération proposée est basée sur l'algorithme de Metropolis-Hastings et donne une distribution uniforme sur $\mathcal{C}(\mathcal{G})$.

Comme mentionné ci-dessus, ces méthodes supposent que les graphes sont ordonnés (PEO), ce qui n'est généralement pas le cas. En fait, seuls les graphes dits cordaux peuvent être parfaitement ordonnés [9, Chapitre 4]. Un graphe est cordal s'il ne contient pas de cycle induit de longueur 4 ou plus. Cette approche n'est ainsi utilisable que pour les graphes cordaux, ce qui est une limitation forte.

3.2 Dominance diagonale et orthogonalisation partielle

A notre connaissance, deux méthodes ont été proposées dans la littérature pour générer des matrices de corrélation associées à un graphe donné $\mathcal{G}$ sans nécessiter de structure cordale : la dominance diagonale et l'orthogonalisation partielle.

La dominance diagonale telle que proposée dans [4] peut être utilisée pour générer une matrice PD. L'idée est de construire une matrice symétrique $\tilde{\mathbf{C}}$, où les éléments $\tilde{c}_{ij}$ sont choisis uniformément au hasard dans l'intervalle $[-1,1]$ si $(i,j) \notin \bar{\mathcal{E}}$, et $\tilde{c}_{ij} = 0$ sinon. Ensuite, on applique la transformation

$$\forall i \in \{1,\dots,p\} \quad \text{do } \tilde{c}_{ii} \leftarrow \sum_{\substack{j=1,\dots,p \\ i \neq j}} |\tilde{c}_{ij}| + \text{perturbation positive aléatoire.}$$

Ensuite, on définit $\mathbf{C} = \text{diag}(\tilde{\mathbf{C}})^{-1/2}\, \tilde{\mathbf{C}}\, \text{diag}(\tilde{\mathbf{C}})^{-1/2}$ pour récupérer une matrice dans $\mathcal{C}(\mathcal{G})$.

Une autre méthode est l'orthogonalisation partielle, proposée dans [4]. L'idée est de commencer avec une matrice initiale $\mathbf{C}$ dont les entrées nulles sont incluses dans le motif de parcimonie souhaité $\overline{\mathcal{E}}$. Les arêtes supplémentaires de $\mathbf{C}$ sont ensuite supprimées à l'aide d'un processus d'orthogonalisation partielle basé sur un algorithme de Gram-Schmidt modifié. La procédure repose sur une écriture de $\mathbf{C}$ sous la forme $\mathbf{C} = \mathbf{Q}\mathbf{Q}^\top$, et orthogonalise itérativement chaque ligne $\boldsymbol{q}_{i\cdot}$ par rapport à l'ensemble des lignes $\{\boldsymbol{q}_{j\cdot} s.t. (i,j) \notin \mathcal{E} \text{ et } j < i\}$.

Dans [4], les auteurs suggèrent de commencer par trianguler le graphe $\mathcal{G}$ pour obtenir un graphe cordal, puis d'appliquer la procédure basée sur Cholesky de [3], comme décrit ci-dessus. La matrice obtenue est alors l'initialisation de l'algorithme d'orthogonalisation partielle.

4 Méthode proposée

L'objectif de notre travail est de générer des matrices de corrélation dans $\mathcal{C}(\mathcal{G})$, c'est-à-dire satisfaisant les contraintes (2). De plus, des contraintes supplémentaires peuvent être ajoutées, selon le contexte. Une de nos motivations est de construire des matrices de corrélation avec une distribution qui ressemble à des données du monde réel, notamment en neurosciences. Dans [1], la distribution des corrélations de connectivité cérébrale s'est avérée centrée autour de valeurs positives. Pour refléter cette propriété, nous imposons cette contrainte supplémentaire sur la moyenne : pour $b \geq -1$,

$$\frac{1}{2|\mathcal{E}|} \sum_{i \neq j} c_{ij} \geq b. \tag{3}$$

Prendre $b = -1$ équivaut à n'avoir aucune contrainte.

Nous cherchons à résoudre le problème d'optimisation suivant :

$$\begin{array}{ll} \underset{\mathbf{C}}{\text{minimize}} & \frac{1}{2}\|\mathbf{C} - \bar{\mathbf{C}}\|_F^2, \\ \text{sous réserve de} & \text{contraintes (2) et (3)}, \end{array} \tag{4}$$

avec $\bar{\mathbf{C}}$ une matrice arbitraire donnée. Avec des données réelles, cela peut être la matrice de corrélation empirique.

5 Résultats

Dans nos simulations, nous considérons $p = 51$, $\mathbf{C} \in \mathbb{R}^{51 \times 51}$ et $\bar{\mathbf{C}}$ une matrice de même taille dont les entrées sont tirées d'une distribution uniforme sur l'intervalle $[-1, 1]$. Pour le motif $\mathcal{E}$, nous utilisons un graphe *Erdős-Rényi*. D'autres structures de graphes ont été considérées (*Barabási-Albert*, *Watts-Strogatz*, *Stochastic Block Model*...) mais ne sont pas présentées ici.

5.1 Comparaison avec les autres approches

La Figure 1 montre la densité des éléments non diagonaux et non nuls des matrices de corrélation générées. Nous comparons notre méthode avec deux autres approches : la dominance diagonale et l'orthogonalisation partielle. Nous avons considéré 50 graphes, avec une densité des graphes cibles à 0.5. Dans notre algorithme, la matrice initiale $\bar{\mathbf{C}}$ est obtenues avec des réalisations d'une distribution uniforme sur [-1,1]. Les densités obtenues en utilisant la dominance diagonale sont concentrées autour de valeurs faibles. Notre approche donne des valeurs plus élevées dans la matrice de corrélation que les autres procédures.

5.2 Influence de la contrainte sur la moyenne

Nous examinons maintenant l'impact de la contrainte dans (3), qui modifie le centrage de la distribution des entrées de corrélation. Nous sommes motivés par une application en

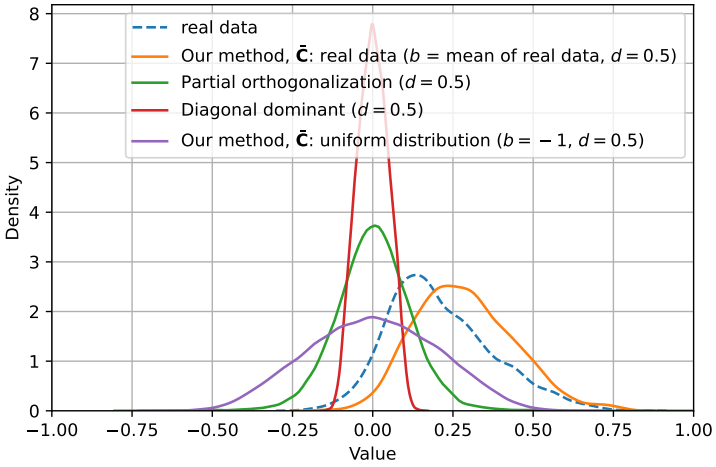


Figure 1: Densité des éléments non diagonaux et non nuls dans les matrices de corrélation générées à l'aide de différentes méthodes.

neurosciences impliquant des données d'IRM fonctionnelle acquises sur des rats. Les données sont accessibles en accès libre et décrites dans [2]. Après le prétraitement expliqué dans [2], des séries temporelles de 51 régions cérébrales pour chaque rat ont été extraites. Nous calculons ensuite la transformée en ondelettes avec l'ondelette de Daubechies d'ordre 8 des 51 signaux. Nous nous intéressons ici à l'échelle d'ondelettes 4, correspondant à l'intervalle de fréquence [0,06 ; 0,12] Hz. Il y a alors 122 coefficients d'ondelettes disponibles pour chacune des 51 régions. Les distributions des corrélations deux à deux entre les coefficients d'ondelettes des régions pour un rat donné. Nous calculons la matrice de corrélation en sélectionnant les 50% d'entrées ayant les valeurs absolues les plus élevées dans la matrice de corrélation empirique, ce qui donne une densité du graphe $d = 50\%$. Les distributions obtenues sont présentées dans la Figure 1.

Nous générons ensuite la matrice synthétique avec notre méthode, avec $d = 50\%$ et le paramètre b égal à la moyenne des entrées dans la matrice de corrélation des données réelles. La matrice initiale $\bar{\mathbf{C}}$ est ici égale à la corrélation empirique des données réelles. La Figure 1 montre que la distribution des données simulées est proche de la distribution réelle.

L'ajout de la contrainte (3) peut entraîner un problème d'optimisation sans solution. La Figure 2 illustre la proportion de cas où une matrice de corrélation valide $\mathbf{C} \in \mathbb{R}^{51\times 51}$ est trouvée pour différentes densités de graphes et valeurs de b dans (4).

Conclusion

La méthode proposée offre plusieurs avantages. Elle ne repose pas sur une structure cordale, garantit la génération d'une matrice semi-définie positive et évite de générer des entrées proches de zéro dans la matrice de corrélatione. De plus, elle permet de générer une matrice de corrélation qui correspond à un graphe lié à une matrice de corrélation empirique.

D'autres analyses, notamment sur l'influence de la structure du graphe sur la distribution

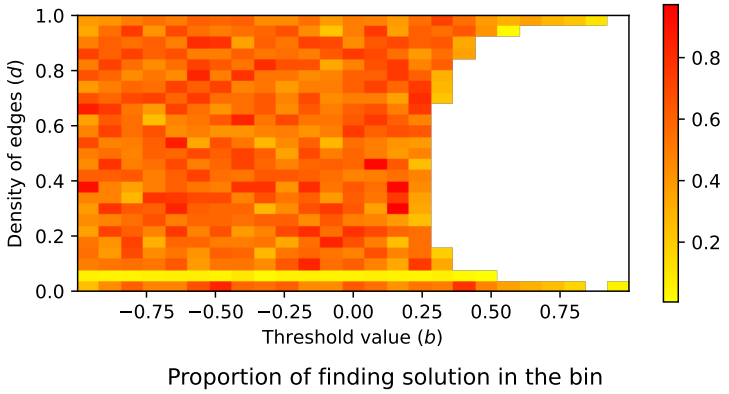


Figure 2: Proportion de réussite à trouver une matrice de corrélation $\mathbf{C} \in \mathbb{R}^{51\times 51}$ en fonction du seuil b, et de la densité du graphe d. Les zones blanches indiquent les cas où aucune solution n'a été trouvée.

des corrélations, sont en cours. Dans la continuité de ce travail, nous souhaitons comparer les performances de divers procédures d'inférence de modèles graphiques sur données simulées.

Bibliographie

[1] Achard, S., Gannaz, I., & Polisano, K. (2022, September). Génération de modèles graphiques. In GRETSI 2022-XXVIIIème Colloque francophone de traitement du signal et des images (pp. 1-3).

[2] Becq, G. J-PC, Habet, T., Collomb, N., Faucher, M., Delon-Martin, C., Coizet, V., Achard, & S. et Barbier, E.L. (2020). Functional connectivity is preserved but reorganized across several anesthetic regimes. *NeuroImage*, 219, 116945.

[3] Córdoba, I., Varando, G., Bielza, C., & Larrañaga, P. (2018). A fast Metropolis-Hastings method for generating random correlation matrices. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 117-124). Cham: Springer International Publishing.

[4] Córdoba, I., Varando, G., Bielza, C., & Larrañaga, P. (2020). On generating random Gaussian graphical models. International Journal of Approximate Reasoning, 125, 240-250.

[5] Horn, R. A., & Johnson, C. R. (2012). Matrix analysis. Cambridge university press.

[6] Joe, H. (2006). Generating random correlation matrices based on partial correlations. Journal of Multivariate Analysis, 97(10), 2177-2189.

[7] Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. Journal of multivariate analysis, 100(9), 1989-2001.

[8] Pourahmadi, M., & Wang, X. (2015). Distribution of random correlation matrices: Hyperspherical parameterization of the Cholesky factor. Statistics & Probability Letters, 106, 5-12.

[9] Vandenberghe, L., & Andersen, M.S. Chordal graphs and semidefinite optimization. (2015). Foundations and Trends in Optimization, 2015, 1(4), 241-433.

Joint estimation of bipartite network collections. Application to plant-pollinator networks

Louis Lacoste [1] & Pierre Barbillon [1] & Sophie Donnet [1]

[1] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France*
{*louis.lacoste,pierre.barbillon,sophie.donnet*}*@agroparistech.fr*

Résumé. On appelle collection de réseaux un ensemble de réseaux ne partageant pas de nœuds mais décrivant le même type d'interactions observées dans différents contextes. Une hypothèse sous-jacente est que ces réseaux partagent une structure commune, puisque la nature des interactions reste la même. Nous proposons d'étendre le modèle à blocs stochastiques bipartites (BiSBM) afin de modéliser conjointement plusieurs réseaux bipartites tout en capturant leur structure de connectivité partagée. Le BiSBM est un modèle probabiliste supposant l'existence de variables latentes représentant des groupes de nœuds (blocs), permettant ainsi une description de la structure du réseau à une échelle mésoscopique. Notre extension, appelée colBiSBM, considère que les réseaux sont des réalisations indépendantes d'un BiSBM commun partageant une structure de connectivité similaire, avec éventuellement des proportions de blocs variables. Nous proposons plusieurs variantes du colBiSBM pour prendre en compte les variations des proportions de blocs en ligne et en colonne entre les réseaux. Les paramètres du modèle et les blocs latents sont estimés à l'aide d'un algorithme EM variationnel, et la sélection du modèle est réalisée à l'aide d'un critère de type BIC adapté. Nous appliquons notre approche à une collection de réseaux d'interactions plantes-pollinisateurs provenant de quatre villes britanniques, montrant que la modélisation conjointe améliore la détection des structures et leur résolution par rapport à des estimations séparées.

Mots-clés. Réseaux bipartites, modèles à blocs stochastiques, modélisation conjointe, réseaux écologiques, interactions plantes-pollinisateurs, algorithme EM variationnel.

Abstract. Let a collection of networks consists of a set of networks that do not share nodes but describe the same type of interactions observed in different contexts. An underlying assumption is that these networks share a common structure since the nature of interactions remains the same. We propose to extend the Bipartite Stochastic Block Model (BiSBM) to jointly model multiple bipartite networks while capturing their shared connectivity structure. The BiSBM is a probabilistic model that assumes the existence of latent variables representing groups of nodes (blocks), allowing for a mesoscopic-scale description of network structures. Our extension, which we call colBiSBM, assumes that the networks are independent realizations of a joint BiSBM that share a common connectivity structure, possibly varying in block proportions. We propose several versions of colBiSBM to account for variations in row and column block proportions across networks. The model parameters and latent blocks are estimated using a variational EM algorithm, and model selection is performed using an ad-hoc Bayesian Information Criterion like. We apply our approach to a collection of plant-pollinator networks from four British cities, demonstrating that joint modelling improves structure detection compared to separate estimations.

Keywords. Bipartite networks, stochastic block models, joint modeling, ecological networks, plant-pollinator interactions, variational EM algorithm.

1 Motivation and Bipartite Stochastic Block Model

Bipartite networks are a common way to represent interactions between two different types of entities. In ecology for instance, they can picture interactions between plants and pollinators. A bipartite network can be represented as a matrix X where the rows and columns represent the two types of entities and the entries are the interactions between them,

$$\begin{cases} X_{ij} = 0 & \text{if no interaction is observed between species } i \text{ and } j \\ X_{ij} \neq 0 & \text{otherwise.} \end{cases}$$

Let us consider a collection of M independent networks where each network is bipartite and has n_1^m row nodes and n_2^m column nodes. We do not assume that the networks involve common nodes. The networks are represented by their incidence matrices $(X^m)_{m\in\{1,\dots,M\}}$. We assume that the networks encode binary observations so $X_{ij} \in \{0,1\}, \forall (i,j)$. $\boldsymbol{X} = (X^1, \dots, X^m)$ denotes the collection of incidence matrices.

Four plant-pollinator networks in British cities We consider a collection of 4 plant-pollinator networks from Baldock et al., 2019. They were collected in Bristol, Edinburgh, Leeds and Reading respectively. The Latent Block Model (LBM) proposed by Govaert and Nadif, 2003 which is a bipartite extension of Stochastic Block Model (in short, BiSBM) is able to capture various complex structures by detecting groups of row nodes and groups of column nodes that share similar connection profiles (see Sander et al., 2015 for instance). The canonical way to treat multiple networks, that we describe next, is to adjust a BiSBM separately on each network.

Separate BiSBM (sep-BiSBM) For network m, let n_1^m (resp. n_2^m) be the number of nodes in row (resp. column) divided into Q_1^m row clusters (resp. Q_2^m column clusters). Let $Z^m = (Z_1^m, \dots, Z_i^m, \dots, Z_{n_1^m}^m)$ and $W^m = (W_1^m, \dots, W_j^m, \dots, W_{n_2^m}^m)$ be independent latent variables such that $Z_i^m = q$ if row node i of network m belongs to row cluster q ($q \in \{1, \dots, Q_1^m\}$) and $W_j^m = r$ if column node j of network m belongs to column block r ($r \in \{1, \dots, Q_2^m\}$). We set

$$\mathbb{P}(Z_i^m = q) = \pi_q^m, \qquad \mathbb{P}(W_j^m = r) = \rho_r^m \tag{1}$$

where $\pi_q^m > 0$, $\rho_r^m > 0$, $\sum_{q=1}^{Q_1^m} \pi_q^m = 1$ and $\sum_{r=1}^{Q_2^m} \rho_r^m = 1$. Given the latent variables Z^m, W^m, the X_{ij}^ms are assumed to be independent and distributed as

$$X_{ij}^m | Z_i^m = q, W_j^m = r \sim \mathcal{B}ern(.; \alpha_{qr}^m) \tag{2}$$

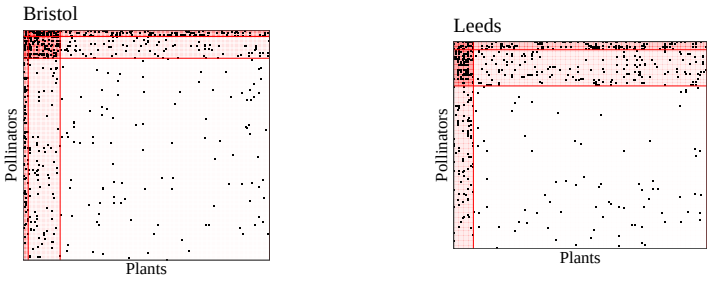


Figure 1: Adjacency matrices of 2 of the 4 British plant-pollinator networks. The nodes being reordered by the marginal distributions in decreasing order. The blocks were obtained by fitting a BiSBM on each network separately. The shades of red show the connectivity parameters $\widehat{\boldsymbol{\alpha}}^m$

where $\mathcal{B}ern$ is the Bernoulli distribution for binary interactions. Equations (1) and (2) define the BiSBM model and we will now use a short notation:

$$\forall m \in \{1, \dots M\},\ X^m \sim \text{BiSBM}_{n_1^m, n_2^m}(Q_1^m, Q_2^m, \boldsymbol{\pi}^{\boldsymbol{m}}, \boldsymbol{\rho}^{\boldsymbol{m}}, \boldsymbol{\alpha}^{\boldsymbol{m}}), \qquad (\textit{sep-BiSBM})$$

where $\boldsymbol{\pi}^m = (\pi_q^m)_{q=1,\dots,Q_1^m}$ and $\boldsymbol{\rho}^m = (\rho_r^m)_{r=1,\dots,Q_2^m}$ are the vectors of the block proportions. The $Q_1^m \times Q_2^m$ matrix $\boldsymbol{\alpha}^m = (\alpha_{qr}^m)_{q=1,\dots,Q_1^m, r=1,\dots,Q_2^m}$ contains the connectivity parameters $\alpha_{qr}^m \in (0,1)$. In this sep-BiSBM model we suppose that each network m follows a BiSBM with its own parameters $(\boldsymbol{\pi}^m, \boldsymbol{\rho}^m, \boldsymbol{\alpha}^m)$.

sep-BiSBM on the British plant-pollinator networks We fit the sep-BiSBM on our 4 networks with the `R` package `sbm` (Chiquet et al., 2024; Leger et al., 2021). The package implements a variational estimation of the parameters and perform model selection for the number of clusters Q_1^m and Q_2^m with an ICL criterion.

We obtain $\widehat{Q_1^1} = \widehat{Q_1^2} = \widehat{Q_1^3} = \widehat{Q_1^4} = 3$ for all networks and $\widehat{Q_2^1} = \widehat{Q_2^2} = \widehat{Q_2^4} = 3$ for Bristol, Edinburgh and Reading, and $\widehat{Q_2^3} = 2$ for Leeds. In Figure 1, we observe a structure commonly found in plant-pollinator networks, that is nestedness, i.e, specialist species (with few connections) interacting mostly with generalist species (with many connections). We only show Bristol as the structures found for Edinburgh and Reading are similar. But Leeds shows a simpler structure with only 2 column blocks. Seeing the similarities in the structures of the networks and their context suggests that a joint modelisation of the networks could be beneficial.

2 Joint modelisation of multiple bipartite networks

We propose several joint probabilistic models that offer to share a connectivity structure between all networks in a collection of bipartite networks. The $(X^m)_{m=1,\dots,M}$ are still assumed to be independent.

A collection of i.i.d. bipartite SBM This first model assumes that all the networks are independent realizations of the same Q_1-Q_2-BiSBM with identical parameters. The *iid*-colBiSBM is defined as follows:

$$\forall m \in \{1, \dots M\},\ X^m \sim \text{BiSBM}_{n_1^m, n_2^m}(Q_1, Q_2, \boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha}), \qquad (\textit{iid}\text{-colBiSBM})$$

where $\forall (q,r) \in \{1,\dots,Q_1\} \times \{1,\dots,Q_2\},\ \alpha_{qr} \in (0,1),\ \pi_q \in (0,1], \sum_{q=1}^{Q_1} \pi_q = 1$ and $\rho_r \in (0,1], \sum_{r=1}^{Q_2} \rho_r = 1$. In this model, the networks share the same mesoscopic structure encoded in the connectivity matrix $\boldsymbol{\alpha}$ and the various blocks are represented in the same proportions in all the networks.

However, the assumption that block proportions are the same among the networks is a strong assumption. In plant-pollinator networks, the proportion of specialist species can differ between networks and thus the model may benefit from not having the same block proportions but sharing a common connectivity structure. The following models relaxes this assumption on either row, column or both.

Varying block proportions on rows π-colBiSBM model still assumes that the networks share a common connectivity structure represented by $\boldsymbol{\alpha}$ but that each network has its own row block proportions. For $m \in \{1, \dots, M\}$, the X^m are independent and

$$\forall m \in \{1, \dots M\},\ X^m \sim \text{BiSBM}_{n_1^m, n_2^m}(Q_1, Q_2, \boldsymbol{\pi^m}, \boldsymbol{\rho}, \boldsymbol{\alpha}), \qquad (\pi\text{-colBiSBM})$$

where $\forall (q,r) \in \{1,\dots,Q_1\} \times \{1,\dots,Q_2\},\ \alpha_{qr} \in (0,1),\ \pi_q^m \in [0,1], \sum_{q=1}^{Q_1} \pi_q^m = 1, \forall m \in \{1,\dots,M\}$ and $\rho_r \in (0,1], \sum_{r=1}^{Q_2} \rho_r = 1$.

This model is more flexible than the iid-colBiSBM as it allows the row block proportions to vary between networks and even to be null ($\pi_q^m \in [0,1]$): if $\pi_q^m = 0$ then the block q is not represented in the network m. The connectivity structure is thus a subset of a large connectivity structure common to all networks. As in Chabert-Liddell et al., 2024, to ensure the identifiability of the model we assume that each block q is represented in at least one network, which is equivalent to $\forall q \in \{1,\dots,Q_1\}, \exists m \in 1,\dots,M$ such that $\pi_q^m > 0$. Let us define the support matrix for row blocks $S^{(1)}$ of size $M \times Q_1$ such that $\forall (m,q),\ S_{mq}^{(1)} = \mathbb{1}_{\pi_q^m > 0}$.

Then the set of admissible supports is the set of binary matrices with a least one non-null element in each row and column $\mathcal{S}_{Q_1} := \left\{ S^{(1)} \in \mathcal{M}_{M \times Q_1}(\{0,1\}) \mid \forall m,\ \sum_{q=1}^{Q_1} S_{mq}^{(1)} > 0,\ \forall q,\ \sum_{m=1}^{M} S_{mq}^{(1)} > 0 \right\}$. The following models adapt these ideas to columns blocks.

Varying block proportions on columns ρ-colBiSBM model still assumes that the networks share a common connectivity structure represented by $\boldsymbol{\alpha}$ but that each network has its own column block proportions. For $m \in \{1, \dots, M\}$, the X^m are independent and

$$\forall m \in \{1, \dots M\},\ X^m \sim \text{BiSBM}_{n_1^m, n_2^m}(Q_1, Q_2, \boldsymbol{\pi}, \boldsymbol{\rho^m}, \boldsymbol{\alpha}), \qquad (\rho\text{-colBiSBM})$$

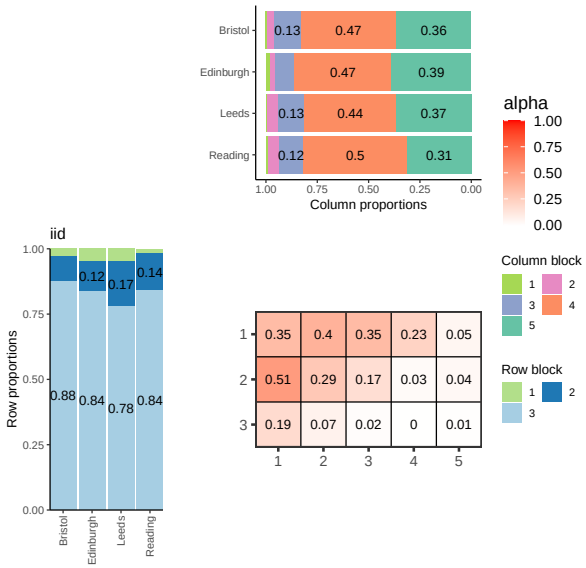


Figure 2: Estimated structure for the *iid*-colBiSBM. The matrix $\widehat{\boldsymbol{\alpha}}$ is shown with red tiles. The block proportions $\widehat{\boldsymbol{\pi}}^m, \widehat{\boldsymbol{\rho}}^m$, for each network are pictured as barplots on the sides. Values are written when above 0.1.

Varying block proportions on rows and columns $\pi\rho$-colBiSBM model still assumes that the networks share a common connectivity structure represented by $\boldsymbol{\alpha}$ but that each network has its own row and column block proportions, it is the least constrained model. For $m \in \{1, \dots, M\}$, the X^m are independent and

$$\forall m \in \{1, \dots M\},\ X^m \sim \text{BiSBM}_{n_1^m, n_2^m}(Q_1, Q_2, \boldsymbol{\pi}^{\boldsymbol{m}}, \boldsymbol{\rho}^{\boldsymbol{m}}, \boldsymbol{\alpha}). \qquad (\pi\rho\text{-colBiSBM})$$

3 Inference and model selection

The log-likelihood being intractable, we use a variational EM algorithm as proposed by Daudin et al., 2008 for SBM. We maximize a variational lower bound of the observed data log-likelihood to estimate the parameters. This is done for a fixed value of (Q_1, Q_2). To find the optimal number of blocks we use a Bayesian Information Criterion Like (BIC-L) as proposed by Chabert-Liddell et al., 2024. And secondly to assess if a joint modelisation is preferred we compare the BIC-L of the sep-BiSBM to the BIC-L of the joint collection. The code is available on Github https://github.com/Chabert-Liddell/colSBM with vignettes.

4 Application to Baldock et al., 2019

In Figure 2 we present the estimated structure $\widehat{\boldsymbol{\alpha}}$ and proportions $\widehat{\boldsymbol{\pi}}$ and $\widehat{\boldsymbol{\rho}}$ for the *iid*-colBiSBM model over the Baldock et al., 2019 networks. The BIC-L favors the 4 joint

models over the sep-BiSBM model and the *iid*-colBiSBM model is preferred over all the other colBiSBM models. A common structure is found and confirms what we assumed looking at the sep-BiSBM. Furthermore, a finer structure is found with this model, i.e. we find 5 column groups. As the proportions of blocks in the four networks are quite similar the *iid*-colBiSBM model is a good fit for the data. The more complex models provide unnecessary degrees of freedom. From an ecological perspective we could then focus on the blocks which are paired and focus on the species that play the same functional role in the 4 networks.

References

Baldock, K. C. R., Goddard, M. A., Hicks, D. M., Kunin, W. E., Mitschunas, N., Morse, H., Osgathorpe, L. M., Potts, S. G., Robertson, K. M., Scott, A. V., Staniczenko, P. P. A., Stone, G. N., Vaughan, I. P., & Memmott, J. (2019). A systems approach reveals urban pollinator hotspots and conservation opportunities. *Nature Ecology & Evolution*, *3*(3), 363–373. https://doi.org/10.1038/s41559-018-0769-y

Govaert, G., & Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, *36*(2), 463–473. https://doi.org/10.1016/S0031-3203(02)00074-2

Sander, E. L., Wootton, J. T., & Allesina, S. (2015). What Can Interaction Webs Tell Us About Species Roles? *PLoS Computational Biology*, *11*(7), e1004330. https://doi.org/10.1371/journal.pcbi.1004330

Chiquet, J., Donnet, S., & Barbillon, P. (2024, September 16). *Sbm: Stochastic Blockmodels* (Version 0.4.7). Retrieved November 4, 2024, from https://cran.r-project.org/web/packages/sbm/index.html

Leger, J.-B., Barbillon, P., & Chiquet, J. (2021, December 1). *Blockmodels: Latent and Stochastic Block Model Estimation by a 'V-EM' Algorithm* (Version 1.1.5). Retrieved November 4, 2024, from https://cran.r-project.org/web/packages/blockmodels/index.html

Chabert-Liddell, S.-C., Barbillon, P., & Donnet, S. (2024). Learning common structures in a collection of networks. An application to food webs. *The Annals of Applied Statistics*, *18*(2), 1213–1235. https://doi.org/10.1214/23-AOAS1831

Daudin, J.-J., Picard, F., & Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, *18*(2), 173–183. https://doi.org/10.1007/s11222-007-9046-7

IA générative - Modèles de diffusion

When Are Two Scores Better Than One? Investigating Ensembles of Diffusion Models

Raphaël Razafindralambo[1,3],
Rémy Sun[1], Frédéric Precioso[1], Damien Garreau[2],
& Pierre-Alexandre Mattei [1]

[1] *Université Côte d'Azur, Inria, Maasai, LJAD/I3S, CNRS*
[2] *Julius-Maximilians-Universität Würzburg*
[3] *raphael.razafindralambo@inria.fr*

Résumé. Les modèles de diffusion sont devenus des méthodes de référence pour générer des données de haute qualité et diversifiées. Par ailleurs, l'attention en apprentissage automatique se porte de plus en plus sur la conception de modèles encore plus performants. En apprentissage supervisé, la combinaison de plusieurs modèles à l'aide d'ensembles est une approche bien établie pour améliorer les performances. Cependant, leur application aux modèles de diffusion basés sur le score, en particulier pour la génération inconditionnelle d'images, reste largement inexplorée. Dans ce travail, nous étudions les bénéfices des méthodes d'ensemble en proposant et en évaluant des stratégies d'ensembles faciles à mettre en œuvre sur le jeu de données CIFAR-10. De plus, nous montrons dans un théorème que K modèles réduit la perte de score-matching.

Mots-clés. Méthodes d'ensemble, modèles de diffusion, score matching

Abstract. Diffusion models have become leading methods for generating high-quality and diverse data. Moreover, the focus in machine learning shifts toward building even more powerful models. In supervised learning, combining multiple models through ensembles is a well-established approach to improve performance. However, their application to score-based diffusion models, particularly for unconditional image generation, remains largely unexplored. In this work, we investigate the benefits of ensemble methods by proposing and evaluating an easy-to-implement ensemble strategy on CIFAR-10 dataset. Moreover, we analytically show that K models reduces the score-matching loss.

Keywords. Ensemble methods, diffusion models, score matching

1 Introduction

Les modèles de diffusion sont une classe performante de modèles génératifs, utilisés notamment en génération d'images [4], en synthèse texte-image [14, 15, 16] et en génération vidéo [10, 19]. Basés sur des processus stochastiques, ils transforment progressivement des données complexes en bruit gaussien via un processus direct, puis apprennent à inverser cette transformation à l'aide d'un réseau de neurones [9, 17] qui estime le score d'Hyvärinen [11]. Bien que

performants, ces modèles restent coûteux en entraînement. Dans ce papier, nous abordons une approche pour améliorer leur efficacité qui est l'utilisation des méthodes d'ensembles, bien connues en apprentissage automatique pour leur capacité à combiner plusieurs modèles afin d'obtenir de meilleures résultats. Ces méthodes ont été adaptées aux réseaux profonds par exemple via le dropout [18, 6] et les ensembles profonds [8, 13]. Ainsi, nous explorons l'application des méthodes d'ensembles aux modèles de diffusion. Pour ce faire nous vérifions, par une analyse théorique et empirique, que assembler K modèles de score améliore l'estimation de celui-ci et la vraisemblance du modèle.

2 Fondements des modèles de diffusion

Les modèles de diffusion constituent une classe de modèles génératifs qui apprennent à transformer progressivement un bruit aléatoire en une donnée structurée de distribution $q_0(\mathbf{x})$ à travers une séquence d'étapes successives [17]. Ce processus repose sur une équation de diffusion dite *forward*, définie le long de $[0, T]$ par $\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$, où $\mathbf{f} : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d$ et $g : [0, T] \to [0, +\infty)$ sont pré-assignées tels que si $\{\mathbf{x}(t)\}_{t=0}^T$ est solution et $\mathbf{x}_0 = \mathbf{x}(0) \sim q_0$, alors $\mathbf{x}_T = \mathbf{x}(T) \sim q_T \approx \mathcal{N}(\mathbf{0}, \sigma_T^2 \boldsymbol{I})$ où $\sigma_T > 0$. Ce processus de diffusion, qui est donc un processus de bruitage, peut être inversé via une équation dite *backward* [1], qui s'écrit

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log q_t(\mathbf{x})\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}, \tag{1}$$

où $q_t(\mathbf{x})$ est la distribution marginale de $\mathbf{x} = \mathbf{x}(t)$ au temps t, le terme $\nabla_{\mathbf{x}} \log q_t(\mathbf{x})$ correspond à son score, et $\bar{\mathbf{w}}$ est un processus de Wiener dans la direction opposée. On souhaite obtenir $\mathbf{x}_0 \sim q_0$ en partant de $\mathbf{x}_T \sim q_T$ et en résolvant cette équation différentielle stochastique pas à pas. L'estimation de $\nabla_{\mathbf{x}} \log q_t(\mathbf{x})$ étant intractable, elle est approchée par une fonction paramétrique $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ minimisant l'objectif de score matching

$$L_{\mathrm{DSM}}(\boldsymbol{\theta}) = \mathbb{E}_{t\sim[0,T]} \left(\lambda(t)\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{\mathbf{x}_t|\mathbf{x}_0}\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)\|^2\right) \tag{2}$$

où $\lambda(t) > 0$, $\mathbf{x}_0 \sim q_0$, $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ est le noyau de transition gaussien et $\log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ est tractable. Associé à l'équation 1, il existe un processus déterministe [1] qui partage les mêmes marginales pour tout t, qui est

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \tilde{\mathbf{f}}(\mathbf{x}, t) := \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log q_t(\mathbf{x})\right]. \tag{3}$$

Dans ce cadre précis, si l'on remplace le score $\nabla_{\mathbf{x}} \log q_t(\mathbf{x})$ par $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$, la trajectoire résultante $\{\mathbf{x}_{\boldsymbol{\theta}}(t)\}_{t=0}^T$ permet d'évaluer explicitement la vraisemblance du modèle $p_{0,\boldsymbol{\theta}}(\mathbf{x})$ qui est sa loi marginale à l'instant initial, à partir de la valeur à l'instant T via une formule de changement de variable instantané [3]. Nous utilisons cette astuce pour nos expériences dans la Section 5.

3 Travaux connexes

Bien que les méthodes d'ensembles soient largement utilisées dans l'apprentissage supervisé, la recherche sur leur application aux modèles de diffusion reste relativement inexplorée. L'utilisation des modèles de diffusion pour l'estimation d'incertitude a été étudiée en employant des hyper-réseaux [7] afin de créer un ensemble de modèles et de combiner les prédictions

finales du processus de diffusion [2]. Une méthode efficace génère un ensemble de modèles de diffusion via les ponts de Schrödinger, évitant des passes multiples [12]. Dernièrement, le PoE au sens de la composition de modèles conditionnels a été étudié pour les modèles de diffusion [5]. Cependant, bien que ces travaux étudient des ensembles de modèles de diffusion, aucun d'eux n'évalue la performance intrinsèque de l'agrégation des prédictions des scores, en comparaison à un modèle unique, notamment en termes d'estimation de densité, d'adéquation du score, dans un cadre inconditionnel. De plus, à notre connaissance, aucun travail antérieur n'a proposé une analyse théorique similaire au nôtre pour les ensembles de modèles de diffusion.

4 K modèles estiment mieux le score que 1 modèle

Considérons un ensemble de taille K, composé de K modèles de score dépendant du temps $\mathbf{s}_{\boldsymbol{\theta}}^{(1)}(\mathbf{x}, t), \ldots, \mathbf{s}_{\boldsymbol{\theta}}^{(K)}(\mathbf{x}, t)$, chacun étant entraîné à prédire le véritable score $\nabla_{\mathbf{x}} \log q_t(\mathbf{x})$ à chaque t. On suppose que ces modèles sont indépendants et identiquement distribués (i.i.d). Nous considérons alors à chaque instant t la moyenne de ces fonctions notée $\overline{s}_{\boldsymbol{\theta}}^{(K)}(\mathbf{x}, t)$. En nous appuyant sur ces hypothèses, nous établissons un théorème montrant que l'agrégation des modèles permet de réduire la perte moyenne de score matching.

Théorème 4.1 *Soient $s_{\boldsymbol{\theta}}^{(1)}(\mathbf{x}, t), \ldots, s_{\boldsymbol{\theta}}^{(K)}(\mathbf{x}, t)$ $K > 0$ modèles de score i.i.d suivant ν construits pour prédire le score. Notons respectivement $L_{DSM}^{(k)}(\boldsymbol{\theta})$ et $\overline{L}_{DSM}^{(K)}$ l'objectif de score matching associé au modèle k et à la moyenne. Alors,*

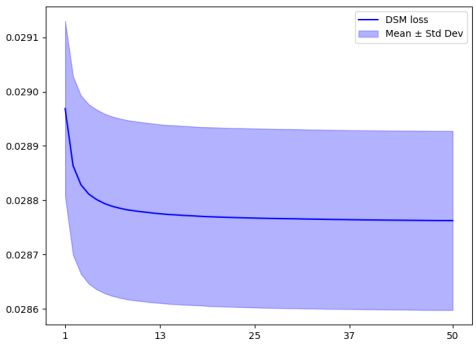
$$\overline{L}_{DSM}^{(K)}(\boldsymbol{\theta}) \leq \frac{1}{K} \sum_{k=1}^{K} L_{DSM}^{(k)}(\boldsymbol{\theta}) \quad \textit{et} \quad \mathbb{E}_{\nu}[\overline{L}_{DSM}^{(K)}(\boldsymbol{\theta})] \leq \mathbb{E}_{\nu}[L_{DSM}^{(1)}(\boldsymbol{\theta})] \tag{4}$$

5 Résultats expérimentaux sur CIFAR-10

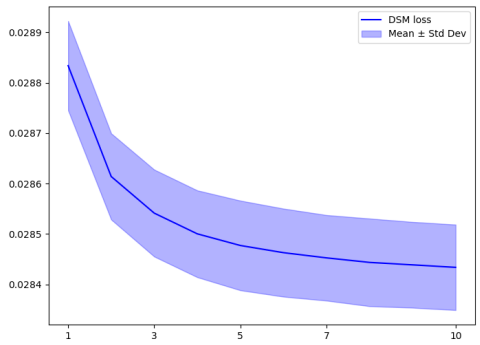
Nous évaluons empiriquement les ensembles sur les images issus de CIFAR-10 en mesurant l'objectif de score-matching, la log-vraisemblance négative (NLL), et la qualité d'image, sur plusieurs méthodes d'ensembles. Pour créer K modèles échangeables, nous fabriquons notre ensemble de modèles en utilisant le MC dropout [6], les ensembles profonds [13]. En terme de méthodes d'aggrégation, nous considérons la moyenne arithmétique. Enfin, pour ces dernières, nous employons l'architecture DDPM++ cont. et une diffusion de type Variance Preserving [17], qui sont entraînés avec le même nombre d'étapes et en utilisant les mêmes paramètres (learning rate, dropout, ...) que le papier d'origine.

6 Conclusion

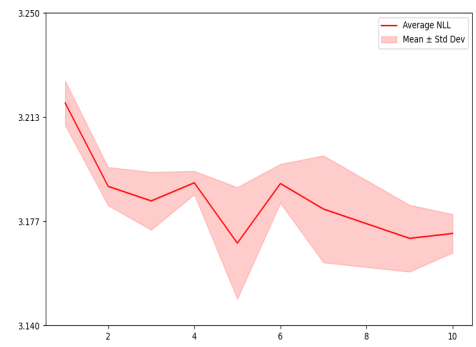
Combiner plusieurs modèles affine l'estimation du score et permet d'augmenter la vraisemblance du modèle. Pour un modèle génératif, maximiser la vraisemblance peut avoir plusieurs applications, comme la compression d'image sans perte.



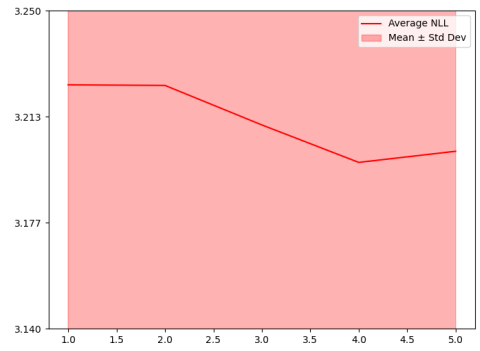
(a) L_{DSM} - MC Dropout



(b) L_{DSM} - Deep Ensemble



(c) NLL (en bits par dimension) - MC Dropout



(d) NLL (en bits par dimension) - Deep Ensemble

Figure 1: Comparaison des méthodes MC Dropout et Deep Ensemble en termes de perte de score-matching et NLL en fonction de la taille K de l'ensemble en abscisse. Chaque expérience est répétée 3 fois.

References

[1] Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[2] Matthew A Chan, Maria J Molina, and Christopher A Metzler. Hyper-diffusion: Estimating epistemic and aleatoric uncertainty with a single model. *arXiv preprint arXiv:2402.03478*, 2024.

[3] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[5] Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: compositional generation with energy-based diffusion models and MCMC. In *International conference on machine learning*, volume 202, pages 8489–8510. PMLR, 2023.

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning*, volume 48, pages 1050–1059. PMLR, 2016.

[7] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

[8] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[10] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[11] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[12] Hyunsu Kim, Jongmin Yoon, and Juho Lee. Fast ensembling with diffusion schr\" odinger bridge. *arXiv preprint arXiv:2404.15814*, 2024.

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30:6402–6413, 2017.

[14] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[16] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[17] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR*, 2021.

[18] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[19] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022.

Suppose that we want to sample from a certain probability density p on $\mathbb{R}^d$. One way would be to sample a point x_0 from an easy-to-sample distribution μ_0 like $\mathcal{N}(0, \mathrm{Id})$ and then apply an algorithm $\mathbb{A}$ to x_0 such that the law of $\mathbb{A}(x_0)$ is p when $x_0 \sim \mu_0$. Score-Based Diffusion Models and Probability Flow ODE (Song et al., 2020) choose an algorithm $\mathbb{A}$ that approximates the flow of probability measures

$$\mu_t \overset{d}{=} tX + \sqrt{1-t^2}Y, \quad t \in [0,1], \tag{1}$$

with $X \sim p$ and $Y \sim \mathcal{N}(0, \mathrm{Id})$ independent random variables. In order to simulate μ_t these methods use two different formulations:

- **The stochastic approach (Score-Based Diffusion):** The Ornstein–Uhlenbeck process initialized at a certain density p

  $$dX_t = -X_t dt + \sqrt{2} dB_t, \quad X_0 \sim p,$$

  satisfies that

  $$X_t \overset{d}{=} \mu_{e^{-t}}.$$

  Then, taking $\tau > 0$, and reverting the process with the following SDE:

  $$dY_t = \big(Y_t + 2\nabla \log \mu_{e^{-(\tau-t)}}(Y_t)\big)dt + \sqrt{2} dB_t, \quad Y_0 \sim \mathrm{law}(X_\tau), \tag{2}$$

  it can be shown that under minimal conditions, the final time Y_τ satisfies $Y_\tau \overset{d}{=} p$ (Anderson, 1982). Therefore, one can sample from p by simulating the process $(Y_t)_t$ until time τ.

- **The deterministic approach (Probability flow ODE):** The flow of maps $(x \mapsto X_t(x))_{t\in[0,1)}$ with X_t solution to the equation

  $$\partial_t X_t(x) = \frac{1}{t}\big(X_t(x) + \nabla \log \mu_t(X_t(x))\big), \quad X_0(x) = x, \tag{3}$$

  satisfies that

  $$(X_t)_{\#}\gamma_d \overset{d}{=} \mu_t.$$

  Therefore, one can sample from p by taking $x \sim \gamma_d$ and running the process $X_t(x)$ until time 1.

1 The diffusion and flow estimators

The construction of the estimators follows Oko et al. (2023) and Fukumizu et al. (2024).

1.1 Heuristic

To simulate the processes (2) and (3), the idea is to first approximate the score function

$$s(t,x) := \nabla \log \mu_{e^{-t}}(x). \tag{4}$$

with an estimator $\hat{s}(t,x)$ fitted on data, by minimizing an empirical version of the *score-matching loss*

$$\int_{t_0}^{\tau} \int_{\mathbb{R}^d} \|\hat{s}(t,x) - s(t,x)\|^2 \mu(t,x)\, dx\, dt, \tag{5}$$

for $0 < t_0 < \tau$. Then using a simple plugin approach:

- **(SDE)** The dynamic (2) is replaced by the process

  $$d\hat{Y}_t = \big(\hat{Y}_t + 2\hat{s}(\tau - t, \hat{Y}_t)\big)dt + \sqrt{2} dB_t, \quad \hat{Y}_0 \sim \mathcal{N}(0, \mathrm{Id}), t \in [0, \tau - t_0]. \tag{6}$$

- **(ODE)** The dynamic (3) is replaced by

  $$\partial_t \hat{X}_t = \frac{1}{t}\big(\hat{X}_t + \hat{s}(\log(t^{-1}), \hat{X}_t)\big), \quad \hat{X}_0 = \mathrm{Id}, t \in [e^{-\tau}, e^{-t_0}]. \tag{7}$$

Note that in practice, the probability flow ODE actually is computed using an approximation $\hat{v}_t$ of the whole vector field $v_t(x) := t^{-1}(x + s(\log(t^{-1}), x))$ by minimizing an empirical version of the *flow-matching loss*

$$\int_{e^{-\tau}}^{e^{-t_0}} \int_{\mathbb{R}^d} \|\hat{v}(t,x) - \frac{1}{t}(x + s(\log(t^{-1}), x))\|^2 \mu(t,x)\, dx\, dt. \tag{8}$$

Although this loss is not directly equivalent to the loss (5) because of the t^{-1} factor, in our setting we estimate the score independently over small intervals $[t_k, t_{k+1}]$ with $t_{k+1} = 2t_k$. As a result, the estimators obtained by minimizing an empirical version of (8) or (5) are effectively equivalent.

1.2 Result

Both methods propose a numerical scheme to simulate this process. Using the regularity results from Stéphanovitch (2024), we show that these methods attain minimax optimal rates of estimation in the Wasserstein distance for a large class of probability densities p.

References

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(82)90051-5. URL https://www.sciencedirect.com/science/article/pii/0304414982900515.

Kenji Fukumizu, Taiji Suzuki, Noboru Isobe, Kazusato Oko, and Masanori Koyama. Flow matching achieves almost minimax optimal convergence, 2024. URL https://arxiv.org/abs/2405.20879.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Arthur Stéphanovitch. Smooth transport map via diffusion process, 2024. URL https://arxiv.org/abs/2411.10235.

Modèles génératifs basés sur le score et échantillonnage de lois conditionnelles

Stanislas Strasman*, Claire Boyer+,‡, Sylvain Le Corff* & Vincent Lemaire*

* *LPSM, Sorbonne Université, UMR CNRS 8001, France,*
stanislas.strasman@sorbonne-universite.fr
+ *LMO, Université Paris-Saclay, UMR CNRS 8628,Orsay, France.*
‡ *Institut Universitaire de France (IUF)*

Résumé. Les modèles génératifs basés sur le score ou modèles de diffusion visent à échantillonner selon une distribution cible en apprenant des fonctions de score (correspondant au gradient du logarithme de densités de probabilité) uniquement à partir d'échantillons de plus en plus bruités de la loi cible (Sohl-Dickstein et al., 2015; Song et al., 2021). Cette classe d'algorithmes s'est imposée comme l'approche privilégiée en apprentissage non supervisé pour l'échantillonnage de loi de probabilités en grande dimension, notamment en synthèse d'images (Dhariwal and Nichol, 2021; Rombach et al., 2022) mais encore en traitement du signal (Chen et al., 2021), en traitement du langage naturel (Hoogeboom et al., 2021), en vision par ordinateur (Yang et al., 2023) et en modélisation multi-modale (Gal et al., 2022; Ruiz et al., 2023).

Cependant, dans de nombreuses applications, il est essentiel d'échantillonner à partir d'une distribution conditionnelle et non seulement de connaître la loi jointe de l'échantillon d'entraînement. C'est par exemple le cas dans les problèmes inverses bayésiens, pour l'imputation de données manquantes ou encore pour la génération conditionnelle sous contrainte d'observations partielles. Afin d'utiliser les modèles de diffusion dans ce cadre, une première approche consiste à incorporer l'information de conditionnement lors de l'apprentissage (Song and Ermon, 2019; Ho et al., 2020). Soit en modifiant le modèle inconditionnel pour qu'il prenne en entrée la variable de conditionnement. Soit en entraînant un modèle conditionnel dédié à prédire ces variables à partir d'entrées partiellement bruitées. Cependant, ces stratégies impliquent de disposer de jeux de données massifs et étiquetés en associant chaque exemple à son information de conditionnement ou bien en concevant puis entraînant de nouveaux modèles spécifiques pour chacune des tâches de conditionnement. Néanmoins, les modèles génératifs sont particulièrement longs et coûteux à entraîner ce qui peut prohiber l'usage de telles techniques. Une alternative plus flexible consiste à utiliser un modèle pré-entraîné et à modifier l'algorithme d'échantillonnage pour générer conditionnellement à partir d'un score appris sur la loi jointe. Dans ce sens des travaux récents ont intégré des outils de filtrage particulaire à la structure diffusive sous-jacente des modèles de score (Wu et al., 2023; Cardoso et al., 2024).

Dans cette optique, nous proposons d'étudier l'intégration de méthodes particulaires au sein des schémas de génération des modèles diffusifs visant des lois conditionnelles afin d'améliorer la qualité des échantillons produits. Nous illustrerons l'intérêt de cette approche dans le contexte de l'imputation de données manquantes, où il s'agit de reconstruire des variables non observées à partir d'informations partielles.

Mots-clés. Méthodes génératives, modèles de diffusion, modèles génératifs basés sur le score, lois conditionnelles, échantillonnage, filtre à particules, Monte Carlo séquentielles.

Abstract. Score-based Generative Models, also known as diffusion models, aim to sample from a target distribution by learning score functions (i.e., gradients of log-probability densities) exclusively from increasingly noisy samples of the target distribution (Sohl-Dickstein et al., 2015; Song et al., 2021). This class of algorithms has emerged as the preferred approach in unsupervised learning for high-dimensional probability distribution sampling, notably for image synthesis (Dhariwal and Nichol, 2021; Rombach et al., 2022), as well as in signal processing (Chen et al., 2021), natural language processing (Hoogeboom et al., 2021), computer vision (Yang et al., 2023), and multimodal modeling (Gal et al., 2022; Ruiz et al., 2023).

However, in many applications, it is crucial to sample from a conditional distribution rather than merely knowing the joint distribution. This is the case, for instance, in Bayesian inverse problems, missing data imputation, or conditional generation subject to partial observations. To leverage diffusion models in this setting, one initial approach is to incorporate conditioning information during training (Song and Ermon, 2019; Ho et al., 2020). This can be done either by modifying the unconditional model so that it takes the conditioning variable as input, or by training a dedicated conditional model to predict these variables from partially noisy inputs. Nevertheless, such strategies generally require massive labeled datasets—pairing each example with its conditioning information—or the design and training of new, specialized models for each conditioning task. Yet generative models are particularly time-consuming and expensive to train, which may limit the applicability of these techniques. A more flexible alternative is to use a pretrained model and modify the sampling algorithm to generate conditionally from a learned score on the joint distribution. Along these lines, recent works have integrated particle filtering techniques into the diffusive structure underlying score-based models (Wu et al., 2023; Cardoso et al., 2024).

In this perspective, we propose to study the integration of particle methods into the generative schemes of diffusion models targeting conditional distributions, with the aim of improving the quality of the generated samples. We illustrate the benefits of this approach in the context of missing data imputation, where the goal is to reconstruct unobserved variables from partial information.

Keywords. Generative methods, diffusion models, score-based generative models, conditional distributions, sampling, particle filtering, sequential Monte Carlo.

1 Résumé long

Pour des raisons de clarté de l'exposé supposons que nous avons accès à un échantillon indépendant identiquement distribué de même loi que le vecteur aléatoire $X_0 \in \mathbb{R}^d$ admettant une densité de probabilité $p_0(x)$, où $x \in \mathbb{R}^d$, tel qu'à une permutation près des dimensions $X_0 = \begin{pmatrix} X_0^{\text{obs}} & X_0^{\text{miss}} \end{pmatrix}^\top$. La question devient donc d'échantillonner selon la loi conditionnelle de X_0^{miss} sachant X_0^{obs}.

Cadre théorique des méthodes génératives diffusives

Le principe fondamental repose sur l'utilisation d'un processus markovien gaussien pour transporter une distribution de départ facile à échantillonner vers la distribution cible des données $p_0(x)$. La première étape est une étape de bruitage définie par une équation différentielle stochastique,

$$\mathrm{d}\overrightarrow{X}_t = \sqrt{\frac{\mathrm{d}}{\mathrm{d}t}(\sigma(t)^2)}\mathrm{d}B_t\,, \qquad \overrightarrow{X}_0 \sim p_0$$

avec $\sigma : [0,T] \to \mathbb{R}^+$. Notons p_t la densité de $\overrightarrow{X}_t$ au temps $t \in [0,T]$. La seconde étape, qui est à proprement parler le modèle génératif, consiste à utiliser le processus associé par renversement du temps (Anderson, 1982) tel que $\left(\overleftarrow{X}_t\right)_{t\in[0,T]} = \left(\overrightarrow{X}_{T-t}\right)_{t\in[0,T]}$ avec

$$\mathrm{d}\overleftarrow{X}_t = \frac{\mathrm{d}}{\mathrm{d}t}(\sigma(t)^2)\nabla \log p_{T-t}(\overleftarrow{X}_t)\mathrm{d}t + \sqrt{\frac{\mathrm{d}}{\mathrm{d}t}(\sigma(t)^2)}\mathrm{d}B_t\,, \qquad \overleftarrow{X}_0 \sim p_T\,. \tag{1}$$

Ainsi donc, si la chaîne backward est initialisée en p_T ce processus dit *backward* peut être utilisé pour générer selon la loi cible, car, par construction, $\overleftarrow{X}_T$ a la même distribution que X_0.

Échantillonnage conditionnel par filtre particulaire

Soit $0 = t_0 < t_1 < \ldots < t_N = T$ une discrétisation régulière ($t_{k+1} - t_k = T/N$) de l'intervalle $[0,T]$. Notons la loi jointe du processus *forward* évaluée en des instants discrets

$$\overrightarrow{p}(x^{\mathrm{obs}}_{0:T}, x^{\mathrm{miss}}_{0:T}) := \overrightarrow{p}\left(\overrightarrow{X}^{\mathrm{obs}}_0 = x^{\mathrm{obs}}_{t_0}, \overrightarrow{X}^{\mathrm{miss}}_0 = x^{\mathrm{miss}}_{t_0}, \ldots, \overrightarrow{X}^{\mathrm{obs}}_T = x^{\mathrm{obs}}_{t_N}, \overrightarrow{X}^{\mathrm{miss}}_T = x^{\mathrm{miss}}_{t_N}\right),$$

et pareillement pour le processus *backward*,

$$\overleftarrow{p}(x^{\mathrm{obs}}_{T:0}, x^{\mathrm{miss}}_{T:0}) := \overleftarrow{p}\left(\overleftarrow{X}^{\mathrm{obs}}_0 = x^{\mathrm{obs}}_{t_N}, \overleftarrow{X}^{\mathrm{miss}}_0 = x^{\mathrm{miss}}_{t_N}, \ldots, \overleftarrow{X}^{\mathrm{obs}}_T = x^{\mathrm{obs}}_{t_0}, \overleftarrow{X}^{\mathrm{miss}}_T = x^{\mathrm{miss}}_{t_0}\right).$$

Le processus *backward* $(\overleftarrow{X}_t)_{t\in[0,T]}$ étant markovien, sa densité jointe se factorise comme suit,

$$\overleftarrow{p}_{0:T}(x^{\mathrm{obs}}_{0:T}, x^{\mathrm{miss}}_{0:T}) = \overleftarrow{p}_0\left(x^{\mathrm{obs}}_0, x^{\mathrm{miss}}_0\right) \prod_{t=t_1}^{t_N} \overleftarrow{p}_{t|t-1}\left(x^{\mathrm{obs}}_t, x^{\mathrm{miss}}_t \middle| x^{\mathrm{obs}}_{t-1}, x^{\mathrm{miss}}_{t-1}\right). \tag{2}$$

De plus, si pour tout $k \in \{1, \ldots N\}$, $\overleftarrow{X}^{\mathrm{obs}}_{t_k}$ et $\overleftarrow{X}^{\mathrm{miss}}_{t_k}$ sont indépendants conditionnellement à $\left(\overleftarrow{X}^{\mathrm{obs}}_{t_{k-1}}, \overleftarrow{X}^{\mathrm{miss}}_{t_{k-1}}\right)^\top$ alors la densité jointe du processus *backward* se réécrit comme suit

$$\begin{aligned}\overleftarrow{p}\left(x^{\mathrm{obs}}_{0:T}, x^{\mathrm{miss}}_{0:T}\right) = {} & \overleftarrow{p}_0\left(x^{\mathrm{obs}}_0, x^{\mathrm{miss}}_0\right) \overleftarrow{p}_{1|0}\left(x^{\mathrm{obs}}_1 \middle| x^{\mathrm{obs}}_0, x^{\mathrm{miss}}_0\right) \\ & \prod_{t=t_1}^{t_N} \overleftarrow{p}_{t+1|t}\left(x^{\mathrm{obs}}_{t+1} \middle| x^{\mathrm{obs}}_t, x^{\mathrm{miss}}_t\right) \overleftarrow{p}_{t|t-1}\left(x^{\mathrm{miss}}_t \middle| x^{\mathrm{obs}}_{t-1}, x^{\mathrm{miss}}_{t-1}\right) \\ & \overleftarrow{p}_{T|T-1}\left(\overleftarrow{x}^{\mathrm{miss}}_T \middle| x^{\mathrm{obs}}_{T-1}, x^{\mathrm{miss}}_{T-1}\right).\end{aligned} \tag{3}$$

Cette écriture de la loi jointe permet, en théorie, d'échantillonner séquentiellement selon X_0^{miss} sachant $X_{0:T}^{\text{obs}}$, dans la mesure où la densité de probabilité $\overleftarrow{p}(x_{0:T}^{\text{miss}}|x_{0:T}^{\text{obs}})$ est proportionnelle à $\overleftarrow{p}(x_{0:T}^{\text{miss}}, x_{0:T}^{\text{obs}})$. De plus, d'après la propriété de *time-reversibility*, pour tout $x_0^{\text{obs}}, x_0^{\text{miss}}, \dots, x_{t_N}^{\text{obs}}, x_{t_N}^{\text{miss}} \in (\mathbb{R}^{d_{\text{o}} \times d_{\text{m}}})^{N+1}$

$$\overrightarrow{p}(x_{0:T}^{\text{obs}}, x_{0:T}^{\text{miss}}) = \overleftarrow{p}(x_{T:0}^{\text{obs}}, x_{T:0}^{\text{miss}}),$$

donc on peut facilement avoir accès à des échantillons de $\overleftarrow{X}_{0:T}^{\text{obs}}$ en utilisant le processus *forward* $\overrightarrow{X}_{0:T}^{\text{obs}}$. La factorisation (3) a inspiré l'algorithme 1 qui intègre la mécanique *bootstrap* de filtrage particulaire aux modèles de diffusion. Enfin, il apparaît en utilisant la propriété de *time-reversibility* que $\overleftarrow{p}\left(x_T^{\text{miss}} \middle| x_{0:T}^{\text{obs}}\right) = p(x_0^{\text{miss}}|x_0^{\text{obs}})$. Il suit que, modulo l'erreur de Monte Carlo, l'algorithme 1 échantillonne bien selon la loi de X_0^{miss} sachant X_0^{obs}.

Mise en pratique

La section précédente étudiait le modèle génératif issu de l'équation (1). Néanmoins, ce dernier ne peut jamais être mis en place en pratique tel quel, pour trois raisons :

- la **loi de p_T n'est pas connue** car il s'agit d'un produit de convolution dépendant de p_0. En pratique, la chaîne backward est initialisée selon une loi gaussienne isotrope p_∞ proche de p_T.
- le **score joint** $\nabla \log p_{T-t}(\overleftarrow{X}_t^{\text{obs}}, \overleftarrow{X}_t^{\text{miss}})$ est rarement connu explicitement, le plus souvent il est **approché** par un réseau de neurones profond $s_\theta : [0,T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$, entraîné à minimiser le score conditionnel (Vincent, 2011).
- les noyaux de transition exacts du processus *backward* ne sont pas connus, il faut **discrétiser** la dynamique continue.

En prenant en compte ces trois erreurs ainsi que l'erreur de Monte Carlo associée à l'algorithme 1 nous nous proposons de dériver une borne d'erreur d'un algorithme déployable en pratique pour la simulation de loi conditionnelle et de tirer parti de celle-ci afin d'optimiser la qualité de génération des données conditionnelles.

Références

B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3) :313–326, 1982.

Gabriel Cardoso, Yazid Janati El Idrissi, Sylvain Le Corff, and Eric Moulines. Monte carlo guided denoising diffusion models for bayesian linear inverse problems. In *International Conference on Learning Representations*, 2024.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad : Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors,

Algorithm 1 Filtrage *bootstrap* pour les diffusions

Input :
  K : nombre de particules.
  N : nombre de pas de discrétisation tels que $t_0 = 0 < \ldots < t_N = T$.
  x^{obs} : le vecteur des dimensions observées.
  $\overrightarrow{x}^{\text{obs}}_{t_k \in \{t_1, \ldots, t_{N-1}\}}$ une trajectoire d'observations bruitées à partir de x^{obs}.
for $i \leftarrow 1$ to K **do**
  Échantillonnage de $\overleftarrow{X}^{(i)}_{t_0} \sim \mathcal{N}(0, \sigma^2(T) I_d)$.
  Assignation des poids initiaux :

$$\tilde{\omega}_0^{(i)} = \overleftarrow{p}_{t_1|t_0}(\overrightarrow{x}^{\text{obs}}_{t_{N-1}} | \overleftarrow{x}^{\text{obs},(i)}_{t_0}, \overleftarrow{x}^{\text{miss},(i)}_{t_0})$$

end for
for $k \leftarrow 1$ to N **do**
  for $i \leftarrow 1$ to K **do**
    Propagation des particules :

$$\tilde{X}^{\text{miss},(i)}_{t_k} \sim \overleftarrow{p}_{t_k|t_{k-1}}(\tilde{X}^{\text{miss},(i)}_{t_k} | X^{\text{obs},(i)}_{t_{k-1}}, X^{\text{miss},(i)}_{t_{k-1}})$$

    if $k < N$:
    Calcul des poids :

$$\tilde{\omega}_k^{(i)} = \tilde{\omega}_{k-1}^{(i)} \overleftarrow{p}_{t_k|t_{k-1}}(\overleftarrow{x}^{\text{obs}}_{t_{N-k-1}} | \overleftarrow{x}^{\text{obs},(i)}_{t_k}, \overleftarrow{x}^{\text{miss},(i)}_{t_k})$$

  end for
  Normalisation des poids :

$$\omega_k^{(i)} = \frac{\tilde{\omega}_k^{(i)}}{\sum_{j=1}^K \tilde{\omega}_k^{(j)}}, \quad i = 1, \ldots, K.$$

end for
Return : x^{obs}, $\tilde{X}^{\text{miss},(i)}_{t_k}$ and $\omega_{N-1}^{(i)}$.

Advances in Neural Information Processing Systems, pages 8780–8794. Curran Associates, Inc., 2021.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word : Personalizing text-to-image generation using textual inversion, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion : Learning categorical distributions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12454–12465. Curran Associates, Inc., 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth : Fine tuning text-to-image diffusion models for subject-driven generation, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11918–11930, 2019.

Yee Whye Song, Jascha Sohl-Dickstein, Durk P Kingma, Avinash Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7) :1661–1674, 2011. doi : 10.1162/NECO_a_00142.

Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models : A comprehensive survey of methods and applications, 2023.

Taking a Big Step: Large Learning Rates in Denoising Score Matching Prevent Memorization

Yu-Han Wu[1], Pierre Marion[2], Gérard Biau[3] & Claire Boyer[4]

[1] *LPSM, Sorbonne Université,* *yu-han.wu@etu.sorbonne-universite.fr*
[2] *Institute of Mathematics, EPFL,* *pierre.marion@epfl.ch*
[3] *LPSM, Sorbonne Université,* *gerard.biau@sorbonne-universite.fr*
[4] *LMO, Université Paris-Saclay,* *claire.boyer@universite-paris-saclay.fr*

Abstract. Denoising score matching plays a pivotal role in the performance of diffusion-based generative models. However, the empirical optimal score–the exact solution to the denoising score matching–leads to memorization, where generated samples replicate the training data. Yet, in practice, only a moderate degree of memorization is observed, even without explicit regularization. In this paper, we investigate this phenomenon by uncovering an implicit regularization mechanism driven by large learning rates. Specifically, we show that in the small-noise regime, the empirical optimal score exhibits high irregularity. We then prove that, when trained by stochastic gradient descent with a large enough learning rate, neural networks cannot stably converge to a local minimum with arbitrarily small excess risk. Consequently, the learned score cannot be arbitrarily close to the empirical optimal score, thereby mitigating memorization. To make the analysis tractable, we consider one-dimensional data and two-layer neural networks. Experiments validate the crucial role of the learning rate in preventing memorization, even beyond the one-dimensional setting.

Keywords. diffusion models, denoising score matching, implicit regularization, neural networks

1 Introduction

Diffusion models have achieved remarkable success in generative modeling across a wide range of tasks, including computer vision, temporal data modeling, multimodal modeling, and natural language processing. Using diffusion models for generative modeling was first proposed by. However, despite the remarkable effectiveness of denoising score matching, the theoretical properties of this training procedure remain unclear. In particular, if the score matching objective were solved perfectly, meaning that the learned score is equal to the score of the empirical data distribution, then the distribution generated by the diffusion process would exactly coincide with the empirical distribution (Li et al., 2024), a failure mode known as *memorization*. Avoiding this issue is crucial in terms of privacy, intellectual property rights, and ability of models to effectively generate new, unseen data.

Empirically, only a moderate amount of memorization is observed in practical settings, even without explicitly regularizing the training objective. This suggests the existence of an

implicit regularization mechanism that prevents exact solving of the denoising score matching problem and thus full memorization of the training data. However, the nature of this regularization is an open problem, as already highlighted in Biroli et al. (2024).

In our work, we approach the question through the lens of the (ir)regularity of the empirical optimal score, in connection with the learning rate of stochastic gradient descent (SGD) (Mulayoff et al., 2021; Qiao et al., 2024). We first show that the empirical optimal score is irregular in the sense that its derivative has a large (weighted) total variation. Then, we prove that the empirical optimal score cannot be stably learned by SGD unless the learning rate becomes vanishingly small. Our main result can be informally stated as follows.

Theorem 1. *(informal) Consider the denoising score matching objective $\mathcal{R}_n$ over the class of two-layer neural networks for one-dimensional data. Then, for a sufficiently small level of noise σ and a learning rate $\eta \gtrsim \sigma^2$, the stochastic gradient descent on $\mathcal{R}_n$ cannot stably converge to a local minimum with arbitrarily small excess risk.*

The take-home message is that for σ small enough and η large enough, the learned score cannot be arbitrarily close to the empirical optimal one. To the best of our knowledge, this is the first demonstration of an implicit regularization mechanism in denoising score matching that prevents (full) memorization, thanks to the non-vanishing learning rate used in practice. Focusing on small values of σ is reasonable, as they correspond to the last steps of the backward diffusion, which are known to play a critical role in memorization

2 Diffusion model and denoising score matching

Let p_{true} be an unknown non-atomic distribution on $\mathbb{R}$ with finite variance. Diffusion-based generative models aim to generate new observations following p_{true}, given an i.i.d. sample of p_{true}. The principle is as follows. For $t \in [0, T]$, the forward diffusion

$$d\overrightarrow{X}_t = -\overrightarrow{X}_t dt + \sqrt{2} d\overrightarrow{B}_t, \quad \overrightarrow{X}_0 \sim p_{\text{true}}, \tag{1}$$

can be reversed in time using the backward diffusion

$$d\overleftarrow{X}_t = (\overleftarrow{X}_t + 2\nabla \log p_{T-t}(\overleftarrow{X}_t))dt + \sqrt{2} d\overleftarrow{B}_t, \quad \overleftarrow{X}_0 \sim p_T, \tag{2}$$

where p_t is the probability density of $\overrightarrow{X}_t$, and $\overrightarrow{B}_t$ (resp. $\overleftarrow{B}_t$) is a Brownian motion. Assuming that sampling from p_T is straightforward, the goal is to use the backward equation (2) to generate new observations. However, this requires learning the unknown *score function* $\nabla \log p_t$. To do so, a key observation is that

$$\overrightarrow{X}_t \overset{\mathcal{D}}{=} \mu(t)\overrightarrow{X}_0 + \sigma(t)\xi, \quad \xi \sim \mathcal{N}(0, 1), \tag{3}$$

where $\mu(t) = e^{-t}$ and $\sigma(t) = \sqrt{1 - e^{-2t}}$. In the following, we consider a fixed time t and we drop the time index t to cast the problem in the more general context of denoising. The

denoising score matching objective given observations $x_1, \dots, x_n$ drawn from p_{true} is written as follows

$$\mathcal{R}_n(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\big[(s(Y) + \frac{1}{\sigma^2}(Y-\mu x_i))^2\big]. \tag{4}$$

One can show (Gu et al., 2023; Li et al., 2024) that its minimizer over all measurable L_2 functions is

$$s^\star(y;\mu,\sigma) = \frac{\sum_{i=1}^n (\mu x_i - y) e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}}}{\sigma^2 \sum_{i=1}^n e^{-\frac{(y-\mu x_i)^2}{2\sigma^2}}}, \quad y \in \mathbb{R}. \tag{5}$$

Li et al. (2024) showed that diffusion equipped with $s^\star$ leads to full memorization of the training data. More precisely, consider the backward diffusion

$$d\overleftarrow{X}_t = (\overleftarrow{X}_t + 2s^\star(\overleftarrow{X}_t; \mu(T-t), \sigma(T-t)))dt + \sqrt{2}d\overleftarrow{B}_t, \quad \overleftarrow{X}_0 \sim \mathcal{N}(0,1),$$

run from time $t = 0$ to time $t = T - \delta$. Then the total variation distance between the distribution of $\overleftarrow{X}_{T-\delta}$ and a smoothed version of the empirical measure of the training sample can be made arbitrarily small as $T \to \infty$ and $\delta \to 0$. In practice, however, only a moderate degree of memorization is observed. Our goal is to substantiate this observation by showing that the score fitted with SGD with a large learning rate deviates from $s^\star$, thereby avoiding full memorization.

3 Model, training and stable minimum

Model. In practice, the risk $\mathcal{R}_n$ is optimized over a class of parameterized functions. We consider in the present paper two-layer ReLU networks with m hidden neurons, i.e., a class $\mathcal{S}$ of the form

$$\mathcal{S} = \Big\{ s_\theta : \mathbb{R} \to \mathbb{R} : s_\theta(y) = \frac{1}{m}\sum_{\ell=1}^{m} w_\ell^{(2)} \phi(w_\ell^{(1)} y + b_\ell), w_\ell^{(1)} \in \{\pm 1\}, (w_\ell^{(2)}, b_\ell) \in [-A, A] \times \mathbb{R} \Big\},$$

where $\theta = (w_{1:m}^{(2)}, b_{1:m}) \in \mathbb{R}^{2m}$ and ϕ is the ReLU activation function. At training time, we consider a random initialization of the inner weights $w_\ell^{(1)}$ with values ± 1, keeping them fixed during training. We also constrain the outer weights $w_\ell^{(2)}$ within a ball. This is a mild requirement, one can ensure by taking $A = \frac{C_n}{\sigma^6}$ that the empirical optimal score $s^\star$ is approximated by functions in $\mathcal{S}$. In addition, for simplicity, we denote the risk for s_θ as $\mathcal{R}_n(\theta)$ instead of $\mathcal{R}_n(s_\theta)$.

Training. Denoising score matching is performed by minimizing the objective $\mathcal{R}_n$ using SGD. Since $\mathcal{R}_n(\theta)$ is not directly computable, we use at each iteration j an unbiased estimator, given by

$$\hat{\mathcal{R}}_j(\theta) = \sum_{i\in\mathcal{B}_j} \big(s_\theta(Y_i) + \frac{1}{\sigma^2}(Y_i - \mu x_i)\big)^2,$$

where $\mathcal{B}_j$ is a random subset of $\{1, \dots, n\}$ and $Y_i \sim \mathcal{N}(\mu x_i, \sigma^2)$. Then, SGD updates are obtained as $\theta_{j+1} = \theta_j - m\eta\nabla\hat{\mathcal{R}}_j(\theta_j)$, where $\eta > 0$ denotes the learning rate.

4 Our result

4.1 Study of empirical score function

We first show that the empirical optimal score $s^\star$, defined in (5), becomes irregular for small σ. Without loss of generality, we assume that the sample is ordered, i.e., $x_1 \leqslant \cdots \leqslant x_n$, and let $\Delta = \min_{2\leqslant i\leqslant n}(x_i - x_{i-1})$ denote the minimum spacing between consecutive observations. We then quantify the irregularity of $s^\star$ by the following weighted total variation

$$\mathrm{TV}_\pi^{(1)}(s^\star) = \int_{\mathbb{R}} |s^{\star\prime\prime}(y;\mu,\sigma)|\pi(y;\mu,\sigma)dy.$$

The weight function π is defined by

$$\pi(y;\mu,\sigma) = \begin{cases} \mathbb{E}_{\xi\sim\mathcal{N}(0,\sigma^2)}\big[\min\{\pi^+(y-\xi;\mu,\sigma), \pi^-(y-\xi;\mu,\sigma)\}\big], & \text{if } \mu x_1 \leqslant y \leqslant \mu x_n, \\ 0, & \text{otherwise,} \end{cases}$$

where denoting by $U \sim \mathcal{U}(\{x_1,\ldots,x_n\})$ a uniform draw from the dataset, for $y \in [\mu x_1, \mu x_n]$,

$$\pi^-(y;\mu,\sigma) = \mathbb{P}(\mu U < y)^2\mathbb{E}[y - \mu U \,|\, \mu U < y], \pi^+(y;\mu,\sigma) = \mathbb{P}(\mu U > y)^2\mathbb{E}[\mu U - y \,|\, \mu U > y].$$

Indeed, we have

Theorem 2. *If* $16n^3 e^{-\frac{\mu^2\Delta^2}{4\sigma^2}} \leqslant 1$, *then* $\mathrm{TV}_\pi^{(1)}(s^\star) \geqslant \dfrac{\mu^3 n\Delta^3}{2^{12}\sigma^4}$.

This result can be recast in the diffusion framework (1)–(3). Indeed, in this situation where $\mu = e^{-t}$ and $\sigma = \sqrt{1-e^{-2t}}$, the condition of the theorem is satisfied when t is close to 0. For small t, we have $\frac{\mu^3}{\sigma^4} = \frac{e^{-3t}}{(1-e^{-2t})^2} \geqslant \frac{1}{8t^2}$, and thus

$$\mathrm{TV}_\pi^{(1)}(s^\star) \geqslant \frac{Cn\Delta^3}{8t^2}.$$

4.2 Implicit regularization and memorization

A local minimum $\theta^\star$ of $\mathcal{R}_n$ is said to be *linearly stable* if there exists some $\varepsilon > 0$ such that, for any θ_0 in the ε-ball $\mathcal{B}_\varepsilon(\theta^\star)$, the following condition holds:

$$\limsup_{j\to\infty} \mathbb{E}\|\theta_j - \theta^\star\|_2 \leqslant \varepsilon.$$

Consider a linearly stable local minimum $\theta^\star$ of $\mathcal{R}_n$ the following gives an upper bound on $\mathrm{TV}_\pi^{(1)}(s_{\theta^\star})$,

Proposition 1. *Let* $\theta^\star = (w_{1:m}^{(2)}, b_{1:m}) \in \mathbb{R}^{2m}$. *If* $s_{\theta^\star} \in \mathcal{S}$ *is a linearly stable local minimum of* $\mathcal{R}_n$, *one has*

$$\mathrm{TV}_\pi^{(1)}(s_{\theta^\star}) \leqslant \frac{1}{2\eta} + \frac{\sqrt{\mathcal{R}_n(\theta^\star)}}{2} + \frac{A}{\sqrt{2\pi}\sigma}\max\Big(\sqrt{2n\mathcal{R}_n(\theta^\star)}, \big(\sqrt{2\pi e}A\sigma n\mathcal{R}_n(\theta^\star)\big)^{\frac{1}{3}}\Big).$$

The result together with Theorem 2 give a clue of how large learning rate may prevent $s_{\theta^\star}$ approaching $s^\star$. Indeed, this is what we show in the following

Theorem 3. *Let $\theta^\star \in \mathbb{R}^{2m}$ be a linearly stable local minimum of $\mathcal{R}_n$. Then there exists $\sigma_0 > 0$, depending on μ and the training sample, such that if $\sigma \leqslant \sigma_0$ and $\eta > \frac{2^{12}\sigma^2}{\mu n^2 \Delta}$, one has*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y\sim\mathcal{N}(\mu x_i,\sigma^2)}\left[(s_{\theta^\star}(Y) - s^\star(Y;\mu,\sigma))^2\right] = \mathcal{R}_n(\theta^\star) - \mathcal{R}_n(s^\star) > \frac{\pi n^5 \mu^3 \Delta^3}{2^{36} e^{1/2} A^4 \sigma^4}.$$

The main message is that for a fixed (small enough) σ, if η is sufficiently large, then the weighted L_2 distance between $s^\star$ and $s_{\theta^\star}$ cannot be made arbitrarily small. In the small-noise regime, i.e., when $\sigma \to 0$, the two conditions of Theorem 3 are automatically satisfied. This is true in particular for diffusions (1)–(3) as $t \to 0$. As mentioned in Section 1, the behavior of the diffusion model is more important in this regime when it comes to memorization. Therefore, in this context, Theorem 3 suggests that setting a large learning rate prevents memorization of the training sample.

5 Experiments

We verify our results with numerical experiments. For the experiment, we take our model to be a two-layer ReLU network with 1000 hidden neuron and we scale the number of epochs the model is trained for reversely proportional to the learning rate to ensure comparable convergence across all models. We train the model with various learning rates on 10 training observations sampled from the two-dimensional isostropic Gaussian distribution of deviation 2. We observe that models trained with larger learning rates generate more unseen observations while the smaller learning rates lead to duplication of training observations.

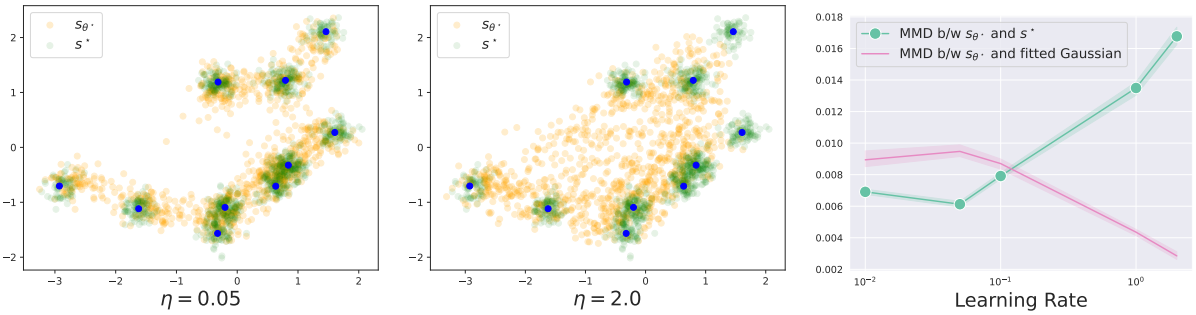


Figure 1: (left) Sample generated by $s^\star$ and $s_{\theta^\star}$ fitted with learning rate 0.05. The training data are the blue points. (middle) Same with $s_{\theta^\star}$ fitted with learning rate 2. (right) The green marked curve corresponds to the MMD between observations generated by $s^\star$ and observations generated by $s_{\theta^\star}$ (for different learning rates). The pink curve is the MMD between observations following the Gaussian distribution fitted on the training data and observations generated by $s_{\theta^\star}$.

6 Conclusion

In our work, we provide a theoretical explanation for why neural networks do not fully memorize the training data. Specifically, our main result, Theorem 3, shows that using a large learning rate acts as a regularization mechanism. This mechanism prevents the network from converging to the empirical optimal score, which becomes unstable under stochastic gradient descent.

Our approach can be extended in several directions. While our theoretical results focus on the one-dimensional case, our experiments strongly suggest that similar results hold in a multivariate setting. Future work could explore this extension, for example by leveraging a multivariate version of the minimum stability framework (Nacson et al., 2023). Next, our results highlight the critical relationship between the learning rate η and the noise variance σ. However, the effects of the sample size n and the dimension d remain unclear. An interesting question is to analyze the role of these hyperparameters, connecting with the literature discussing the impact of n. Finally, even if our results indicate that η should not be too small to prevent memorization, they do not guarantee that a larger learning rate improves the quality of the generated data. Exploring trade-offs between memorization, generation quality, and training speed is a key avenue for future research.

References

G. Biroli, T. Bonnaire, V. de Bortoli, and M. Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

X. Gu, C. Du, T. Pang, C. Li, M. Lin, and Y. Wang. On memorization in diffusion models. *arXiv:2310.02664*, 2023.

S. Li, S. Chen, and Q. Li. A good score does not lead to a good generative model. *arXiv:2401.04856*, 2024.

R. Mulayoff, T. Michaeli, and D. Soudry. The implicit bias of minima stability: A view from function space. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17749–17761. Curran Associates, Inc., 2021.

M. S. Nacson, R. Mulayoff, G. Ongie, T. Michaeli, and D. Soudry. The implicit bias of minima stability in multivariate shallow ReLU networks. In *The Eleventh International Conference on Learning Representations*, 2023.

D. Qiao, K. Zhang, E. Singh, D. Soudry, and Y.-X. Wang. Stable minima cannot overfit in univariate ReLU networks: Generalization by large step sizes. *arXiv:2406.06838*, 2024.

IA générative - VAE

Clustering guidé dans des autoencodeurs variationels

Violaine Courrier [1] & Christophe Biernacki [2]

[1] *Withings, Paris & Inria, Université de Lille ; violaine.courrier@inria.fr*
[2] *Inria, Université de Lille, CNRS, UMR 8524, Lille ; christophe.biernacki@inria.fr*

Résumé. Dans ce travail, nous proposons le Guided Clustering Variational Autoencoder (GCVAE), un modèle qui combine un mélange gaussien et un autoencodeur variationnel. Sa particularité est d'intégrer une "variable guide" pour influencer la formation des clusters, tout en maintenant la possibilité de traiter de nouvelles données où cette variable guide est absente. Le GCVAE exploite à la fois la richesse des informations initiales et les connaissances contextuelles pour une segmentation plus pertinente et plus aisée à interpréter que les méthodes de clustering purement non supervisées.

Mots-clés. Clustering, VAEs, IA générative, inférence variationnelle

Abstract. This article introduces the Guided Clustering Variational Autoencoder (GCVAE), a model that combines a Gaussian Mixture Model (GMM) with a Variational Autoencoder (VAE). Its distinguishing feature is the integration of a "guiding" variable to influence cluster formation, while preserving the ability to handle new data where the guiding variable is unavailable. By doing so, the GCVAE leverages both the richness of the original information and the contextual knowledge, resulting in more relevant and interpretable segmentation than purely unsupervised clustering methods.

Keywords. Clustering, VAEs, Generative AI, variational inference

1 Introduction

Le clustering est une technique couramment employée pour segmenter des données x selon leurs similarités. Pourtant, dans la pratique, des considérations liées au domaine ou à la tâche influencent grandement la formation et l'interprétation des groupes. Par exemple, à partir d'un même jeu de données clients (fréquence d'achats, montants dépensés, engagement, etc.), le service logistique peut classifier ces clients pour optimiser les stocks et anticiper la demande, tandis qu'une équipe marketing privilégiera l'engagement et le comportement d'achat pour adapter ses campagnes promotionnelles.

Les méthodes traditionnelles de clustering ignorent souvent ces facteurs contextuels et se concentrent uniquement sur la séparation ou la variance. Pour pallier cette limite, nous proposons un modèle qui intègre une variable guide y afin d'orienter la formation des clusters tout en préservant la structure initiale des données. Ainsi, notre approche offre une flexibilité

permettant d'adapter le clustering à des objectifs spécifiques (portés par y), sans sacrifier la richesse des informations de départ (portées par x).

Le Guided Clustering Variational Autoencoder (GCVAE) intègre un modèle de mélange gaussien [2] dans un Variational Autoencoder [5]. Il se distingue en générant en plus des données d'entrée x une variable guide y. En utilisant y uniquement dans le modèle génératif, la procédure d'inférence reste dépendante uniquement de x — un choix de conception qui préserve la capacité de regrouper de nouveaux échantillons même lorsque y n'est pas disponible. Ce modèle offre une nouvelle approche du clustering à la fois flexible et interprétable, répondant au besoin de segmentation consciente du contexte dans des jeux de données complexes.

2 Le modèle

Le modèle GCVAE proposé vise à segmenter les données d'entrée $X = (x_1, ..., x_n)$ tout en étant guidé par les informations contextuelles fournies par $Y = (y_1, ..., y_n)$, avec n le nombre d'observations. Nous supposons que les données sont générées par un processus impliquant des variables latentes $Z = (z_1, ..., z_n)$ et des clusters $c \in \{1, ..., K\}$, avec K le nombre prédéfini de clusters. Dans le reste de ce document, on omettra l'indice $i \in \{1, .., n\}$.

2.1 Modèle génératif

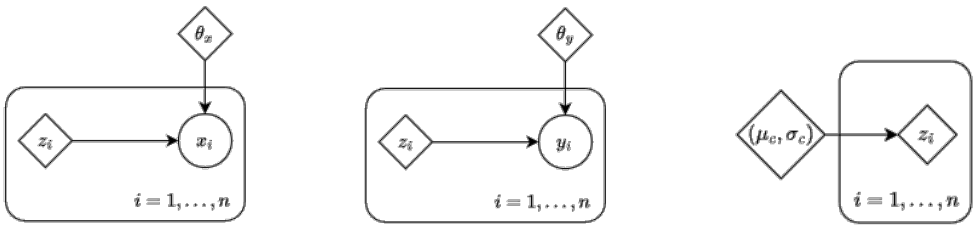


Figure 1: Une représentation graphique des décodeurs dans les Équations 3 et 4, et le GMM dans l'Équation 2, respectivement. *Les losanges indiquent que les variables sont latentes, tandis que les ronds désignent les observations.*

Le processus génératif de notre modèle se compose des étapes suivantes :

1. Choisir un cluster c en utilisant une distribution Multinomiale : $p_{\boldsymbol{\pi}}(c) = Cat(c; \boldsymbol{\pi})$.
2. Générer un vecteur latent z conditionné par le cluster c : $p_{\mu_c, \sigma_c}(z|c) = \mathcal{N}(z; \mu_c, \sigma_c^2 I)$.
3. Générer la variable x à partir du vecteur latent z : $p_{\theta_x}(x|z) = \mathcal{N}(x; f_{\theta_x}(z), I)$.
4. Générer la variable y à partir du vecteur latent z : $p_{\theta_y}(y|z) = \mathcal{N}(y; f_{\theta_y}(z), I)$.

Ici, $\boldsymbol{\pi} = (\pi_1, ..., \pi_K) \in [0, 1]^K$, avec π_c la probabilité a priori pour le cluster c, $\sum_{c=1}^{K} \pi_c = 1$, μ_c et σ_c^2 sont les paramètres de la distribution normale multivariée correspondante au cluster

c, I est la matrice identité, $f_{\theta_a}(z)$ avec $a \in \{x, y\}$ sont des réseaux avec entrée z et paramétrés par θ_a.

Selon le processus génératif ci-dessus, illustré dans la Figure 1, avec $\Theta = \{\mu_c, \sigma_c, \pi_c, \theta_x, \theta_y\}$, la probabilité jointe du modèle est donnée par :

$$p_\Theta(z, c, x, y) = p_\Theta(y|z)p_\Theta(x|z)p_\Theta(z|c)p_\Theta(c). \tag{1}$$

2.2 Modèle d'inférence

Pour effectuer l'inférence, nous devons approcher la loi a posteriori, mais non calculable, $p_\Theta(z, c \mid x, y)$. Nous utilisons alors l'inférence variationnelle [1] pour y parvenir, en utilisant un modèle d'inférence $q_\phi(z, c|x)$ qui repose *uniquement* sur les données d'entrée x. Cette approche garantit que le modèle reste applicable même lorsque la variable guide y n'est pas disponible au moment de la prédiction.

Nous supposons aussi une approximation de champ moyen, pour des raisons de simplicité :

$$q_\phi(z, c|x) = q_\phi(z|x)q(c|x) \tag{2}$$

avec

$$q_\phi(z|x) = \mathcal{N}(z; \tilde{\mu}, \tilde{\sigma}^2 I) \tag{3}$$
$$[\tilde{\mu}, \log \tilde{\sigma}] = g_\phi(x) \tag{4}$$

où $\tilde{\mu} \in \mathbb{R}^J$ et $\tilde{\sigma} \in \mathbb{R}^{J+}$, avec J qui correspond à la dimension de l'espace latent. Le modèle exprimé dans l'Équation 3 est illustré dans la Figure 2. Ici, g_ϕ est un réseau neuronal paramétré par ϕ, et le détail de l'estimation de $q(c|x)$ est décrit à la fin de la partie 2.3.

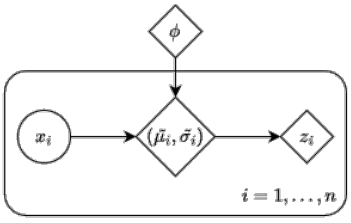


Figure 2: Une représentation graphique du modèle d'inférence.

2.3 Objectif d'entraînement

L'entraînement du GCVAE est guidé par l'objectif d'inférence variationnelle, spécifiquement la *Evidence Lower Bound* (ELBO) que l'on cherche à maximiser :

$$\arg\max_{\Theta,\phi} \text{ELBO}(x, y) = \arg\max_{\Theta,\phi} \mathbb{E}_{q_\phi(z,c|x)} \left[\log \frac{p_\Theta(z, c, x, y)}{q_\phi(z, c|x)} \right].$$

L'ELBO peut être décomposée en plusieurs termes, en utilisant l'Équation 1, avec les termes de reconstruction pour x et y, et les termes de la divergence KL qui régularisent l'espace latent :

$$\text{ELBO}(x, y) = \mathbb{E}_{q_\phi(z|x)}[\log p_\Theta(y|z)] + \mathbb{E}_{q_\phi(z|x)}[\log p_\Theta(x|z)] - \text{KL}[q_\phi(z, c|x)||p_\Theta(z|c)p_\Theta(c)].$$

Elle peut être ensuite développée comme ci-dessous :

$$\begin{aligned}\text{ELBO}(x, y) = &-\frac{1}{L}\sum_{l=1}^{L}||y - f_{\theta_y}(z^{(l)})||_2^2 - \frac{1}{L}\sum_{l=1}^{L}||x - f_{\theta_x}(z^{(l)})||_2^2 \\ &- \frac{1}{2}\sum_{c=1}^{K} q(c|x)\sum_{j=1}^{J}\left(\log\sigma_{cj}^2 + \frac{\tilde{\sigma}_j^2}{\sigma_{cj}^2} + \frac{(\tilde{\mu}_j - \mu_{cj})^2}{\sigma_{cj}^2}\right) \\ &+ \sum_{c=1}^{K} q(c|x)\log\pi_c + \frac{1}{2}\sum_{j=1}^{J}(1 + \log\tilde{\sigma}_j^2) - \sum_{c=1}^{K} q(c|x)\log q(c|x).\end{aligned}$$

avec $z^{(l)} \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2 I)$, et $[\tilde{\mu}, \log\tilde{\sigma}^2] = g_\phi(x)$. L est le nombre d'échantillons Monte Carlo dans l'estimateur *Stochastic Gradient Variational Bayes* (SGVB).

Nous allons approximer $q(c|x)$ en utilisant l'estimateur SGVB :

$$q(c|x) = \mathbb{E}_{q_\phi(z|x)}[p(c|z)] \simeq \frac{1}{L}\sum_{l=1}^{L}\frac{p_\Theta(z^{(l)}|c)p_\Theta(c)}{\sum_{c'} p_\Theta(z^{(l)}|c')p_\Theta(c')}.$$

3 Ajustements du modèle

Pour garantir que le GCVAE capture efficacement l'information fournie par la variable guide y, nous proposons plusieurs ajustements liés à l'architecture du modèle et au processus d'entraînement.

Encodeur et décodeurs simplifiés Pour éviter que l'encodeur et les décodeurs ne capturent directement l'information des groupes, nous optons pour des architectures de réseau plus simples. Nous maintenons également la dimension de l'espace latent petite. Ces décisions aident à garantir la structuration de l'espace latent.

$q(c|x)$ binaire Pour mettre en avant l'apprentissage des paramètres du GMM, nous avons imposé une décision binaire dans la distribution de probabilité $q(c \mid x)$. Cette approche garantit que le modèle attribue chaque individu à un seul cluster, affinant ainsi les frontières des clusters.

Fonction de perte pondérée Inspirés par le cadre beta-VAE [3], nous introduisons des poids dans la fonction de perte pour équilibrer les termes de reconstruction et la structure de l'espace latent, conduisant à une nouvelle formulation de l'ELBO :

$$\begin{aligned}\text{ELBO}_\beta(x,y) = \mathbb{E}_{q_\phi(z|x)}[\log p_\Theta(y|z)] + \mathbb{E}_{q_\phi(z|x)}[\log p_\Theta(x|z)] \\ - \beta \times \text{KL}[q_\phi(z,c|x)||p_\Theta(z|c)p_\Theta(c)].\end{aligned}$$

Nous choisissons $\beta \in \mathbb{R}^+$ de sorte que le terme KL et le terme de reconstruction contribuent de manière similaire à la perte globale. Cette stratégie aide à stabiliser l'entraînement en empêchant un composant unique de dominer le processus d'optimisation.

Initialisation GMM Dans ce travail, une étape de pré-entrainement est utilisée pour initialiser les paramètres du GMM (μ, σ), une pratique courante dans le *deep clustering* [4, 6]. Nous introduisons le poids α dans la fonction de perte pour équilibrer la reconstruction de x, y et la structure de l'espace latent sans les clusters, conduisant à l'ELBO suivante :

$$\begin{aligned}\text{ELBO}_{\alpha,\beta}(x,y) = \alpha \times \mathbb{E}_{q_\phi(z|x)}[\log p_\Theta(y|z)] + \mathbb{E}_{q_\phi(z|x)}[\log p_\Theta(x|z)] \\ - \beta \times \text{KL}[q_\phi(z|x)||\mathcal{N}(z; 0_J, I)]\end{aligned}$$

avec 0_J un vecteur nul de dimension J.
En priorisant la reconstruction de y avec $\alpha > 1$, nous encourageons ainsi le modèle à aligner l'espace latent avec la variable guide.

4 Conclusion

Dans ce travail, nous avons présenté le *Guided Clustering Variational Autoencoder* (GCVAE), qui intègre une variable guide y afin d'orienter la formation des clusters vers des objectifs contextuels. En mariant un modèle de mélange gaussien à un autoencodeur variationnel, GCVAE construit un espace latent interprétable préservant la structure de x tout en exploitant l'information supplémentaire apportée par y. Des expériences présentées pendant la conférence, sur un dataset MNIST-SVHN et un jeu de données réel de l'entreprise Withings [1], confirment que cette approche guidée améliore la pertinence et l'interprétation des clusters.

References

[1] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017.

[1] Withings est une entreprise française qui conçoit des objets médicaux connectés tels que des montres, des balances ou encore des tensiomètres. https://www.withings.com/fr/fr/

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

[3] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[4] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: A generative approach to clustering. *CoRR*, abs/1611.05148, 2016.

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[6] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis, 2016.

VAE MULTI-COMPOSANT AVEC CHAMP DE MARKOV GAUSSIEN : UNE NOUVELLE APPROCHE POUR LA CONCEPTION INDUSTRIELLE COMPLEXE

Fouad Oubari [1,2] & Mohamed El – Baha[2] & Raphael Meunier[2] & Rodrigue Décatoire[2] & Mathilde Mougeot[1,3]
[1] *ENS Paris Saclay,* [2] *Michelin,* [3] *ENSIIE*
{fouad.oubari ; mohamed.el-baha ; raphael.meunier ; rodrigue.decatoire} @michelin.com
mmougeot@ens-paris-saclay.fr

Résumé. Ce travail présente une approche permettant de générer des données issues de systèmes définis par plusieurs composants caractérisés par des dépendances complexes. Ces systèmes sont largement présents dans le milieu industriel. Les approches d'Autoencodeur Variationnel introduites à ce jour dans la littérature reposent sur des schémas d'agrégation simples qui ne parviennent souvent pas à modéliser ces dépendances complexes. Pour surmonter cette limitation, nous proposons une nouvelle architecture d'Autoencodeurs Variationnels qui intègre un champ de Markov aléatoire gaussien dans les distributions a priori et a posteriori, offrant ainsi une meilleure représentation des relations inter-composants. Notre modèle appelé GMRF-MCVAE, atteint des performances de pointe sur un jeu de données simulées par copule, conçu pour représenter la complexité des interactions pouvant être observées en pratique. Il démontre des résultats comparables à l'état de l'art sur le benchmark PolyMNIST. Nous présentons également une application de notre modèle sur un cas industriel d'épures de pneu.

Keywords. Mutli-component datasets, Deep generative models, Variational AutoEncoders

1 Introduction

Les ensembles de données issus de systèmes multi-composants sont omniprésents dans de nombreux domaines d'application. L'élaboration de méthodes capables de modéliser les dépendances des données multi-composants et de générer ces données constitue un enjeu crucial. À titre d'exemple, dans le domaine de l'imagerie médicale, la combinaison de scanners CT et d'IRM pour un même patient fournit des informations distinctes mais complémentaires (Rahimi 2022, Puhr-Westerheide 2019). En dehors du domaine médical, les données multi-composants trouvent également des applications dans des secteurs tels que le multimédia, la finance ou l'industrie (Girdhar 2023, Xie 2024, Cobb 2023).

Plusieurs approches basées sur l'architecture d'Autoencodeurs Variationnels Multi-Composants (MCVAE) introduits dans la littérature, ont démontré de bonnes performances sur des jeux de données multimodales, mais reposent fréquemment sur des schémas d'agrégation simplifiés (Wu 2018, Shi 2019, Sutter 2021, Palumbo 2023). Ces approches sont méthodologiquement limités pour capturer capturer pleinement les dépendances complexes entre les composants.

Les champs de Markov aléatoires (MRF) (koller 2009) sont des outils mathématiques qui offrent naturellement une représentation structurée, en décomposant une distribution en plusieurs facteurs permettant de modéliser explicitement les relations pouvant exister entre les

différents composants d'un système. Cette factorisation permet l'intégration d'interactions plus nuancées dans l'espace latent, ce qui peut aboutir à des générations préservant mieux ces interdépendances.

Afin de surmonter les limites des MCVAEs existants, nous proposons une nouvelle approche, désignée sous le nom de Gaussian Markov Random Field Multi-Component Variational AutoEncoder (GMRF-MCVAE). Cette méthode repose sur l'intégration des champs de Markov aléatoires gaussiens (GMRF) dans les distributions a priori et a posteriori, ce qui améliore la modélisation des interdépendances tout en conservant une inférence efficace.

Les principales contributions de ce travail sont les suivantes :

- La conception d'une approche théorique intégrant les MRF dans les MCVAEs pour capturer des dépendances complexes entre composants.
- L'introduction du GMRF-MCVAE, une architecture exploitant les GMRFs. Nous exposons également les méthodes de génération conditionnelle et inconditionnelle.
- La démonstration des performances état de l'art sur un dataset de copules pour des dépendances complexes et des résultats compétitifs sur le benchmark PolyMNIST.

2 Méthode : GMRF-MCVAE

Notons $X = (x_1 \dots, x_n)$ les jeux de données à n composants fixés. Le GMRF-MCVAE (Figure 1) intègre un champ de Markov aléatoire gaussien dans l'espace latent d'un Autoencodeur Variationnel Multi-Composants. Chaque composant x_i est associé à un encodeur-décodeur : l'encodeur spécifique produit les potentiels unaires ψ_i., tandis qu'un encodeur global génère potentiels par paire $\psi_{i,j}$. Ces potentiels définissent une distribution latente conjointe, capturant à la fois les détails des composants individuels et leurs dépendances.

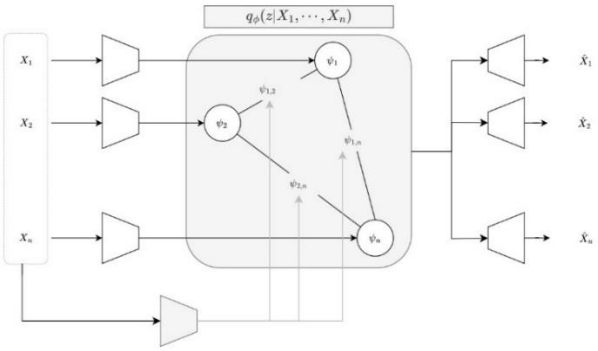


Figure 1 : Schémas du GMRF-MCVAE.

Pour simplifier les calculs et rendre l'échantillonnage différentiable, nous adoptons l'hypothèse que le MRF est gaussien – GMRF. Ce qui revient à définir les potentiels comme suit (Murphy 2012) :

$$\psi_i(z_i) = ex\,p\left(-\frac{1}{2}z_i{}'\Lambda_{i,i}z_i + \eta_i{}'z_i\right), \qquad \psi_{i,j}(z_i, z_j) = exp\,(-\frac{1}{2}z_i'\Lambda_{i,j}z_j)$$

La distribution conjointe du GMRF peut alors s'écrire :

$$p(z) \propto exp\ (\eta' z - \frac{1}{2} z' \Lambda z)$$

Où Λ est la matrice de précision (partitionnée en blocs $\Lambda_{i,j}$) et η définissent les paramètres naturels d'une gaussienne. Cela revient à définir une distribution gaussienne avec

$$p(z) = \mathcal{N}(\mu = \Lambda^{-1}\eta, \Sigma = \Lambda^{-1})$$

Où $\mu = (\mu_1, \dots, \mu_n) \in \mathbb{R}^{d \times n}$ et $\Sigma = \begin{bmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n,1} & \cdots & \Sigma_{n,n} \end{bmatrix}$ avec $\Sigma_{i,j} \in \mathbb{R}^{d \times d}$.

2.1 Échantillonnage inconditionnel

Pour permettre un échantillonnage différentiable, le modèle utilise l'astuce de reparamétrisation (Kingma 2019). Avec la décomposition de Cholesky de la covariance ($\Sigma = LL'$), un échantillon latent est obtenu par :

$$z = \mu + L\varepsilon, o\grave{u}\ \varepsilon \sim N(0, I)$$

2.3 Échantillonnage conditionnel

Dans une distribution gaussienne, l'échantillonnage conditionnel de z_i sachant z_j où $z = (z_1, \dots, z_n) \sim \mathcal{N}(\mu, \Sigma)$ peut s'exprimer analytiquement :

$$p(z_i | z_j) = \mathcal{N}(\tilde{\mu}_i^j, \tilde{\Sigma}_i^j)$$

Avec : $\tilde{\mu}_i^j = \mu_i + \Sigma_{i,j}\Sigma_{j,j}^{-1}(z_j - \mu_j),\ \tilde{\Sigma}_i^j = \Sigma_{i,i} - \Sigma_{i,j}\Sigma_{j,j}^{-1}\Sigma_{i,j}$

Contrairement à d'autres approches MCVAEs (Wu 2018, Shi 2019, Sutter 2021, Palumbo 2023) cette propriété permet de générer des composants conditionnels à partir d'observations partielles, sans nécessiter de réentraînement spécifique du modèle pour cette tâche.

2.2 Construction de la matrice de covariance

La matrice de covariance Σ définie plus haut, est construite sous forme de blocs. Chaque bloc diagonal $\Sigma_{i,i}$ représente les variances spécifiques des composants, générées par les encodeurs individuels, et les blocs hors diagonale $\Sigma_{i,j}$ modélisent les covariances inter-composants, produits par l'encodeur global. Pour garantir que Σ soit symétrique et définie-positif (SPD), nous démontrons le théorème suivant, qui généralise le corollaire classique de Hadamard aux matrices par blocs :

Théorème 1*: Soit Σ une matrice symétrique partitionné par bloc. Si les blocs diagonaux sont SDP et vérifient l'inégalité :*

$$|| \Sigma_{i,i}^{-1} ||^{-1} \geq \sum_{k \neq i} || \Sigma_{i,k} ||^{-1}$$

où //.// désigne la norme spectrale. Alors Σ est définie-positif.

Dans notre modèle, les blocs hors diagonale sont mis à l'échelle de manière adaptative pendant l'entrainement pour respecter les contraintes du théorème 1.

3 Expériences et résultats

Nous évaluons le modèle GMRF-MCVAE en le comparant à quatre approches d'Autoencodeurs Variationnels Multi-Composants : MVAE (Wu 2018) , MMVAE (Shi 2019), MoPoE-VAE (Sutter 2021), et MMVAE+ (Palumbo 2023) sur deux jeux de données :

Copula Dataset : un jeu de données synthétique conçu pour capturer des dépendances complexes entre les composants. Chaque échantillon contient quatre composants bidimensionnels générés à l'aide de copules gaussiennes, reproduisant des relations inter-composants similaires à celles observées dans des domaines comme la finance ou la météorologie (Mackenzie 2014, Tedesco 2023).

PolyMNIST Dataset (Daunhawer 2022)**:** une extension du dataset MNIST où chaque point de données contient cinq images d'un même chiffre avec des styles d'écriture différents et des arrière-plans spécifiques à chaque composant. Ce jeu de données teste la capacité des modèles à préserver une identité commune tout en gérant la complexité des composants.

Résultats :

Sur le Copula Dataset, nous observons que le GMRF-MCVAE produit une distribution générées fidèle à celle des données d'entrainement comme le montre la Figure 2. Ces observations sont confirmées par des distances de Wasserstein significativement plus faibles (Table 1).

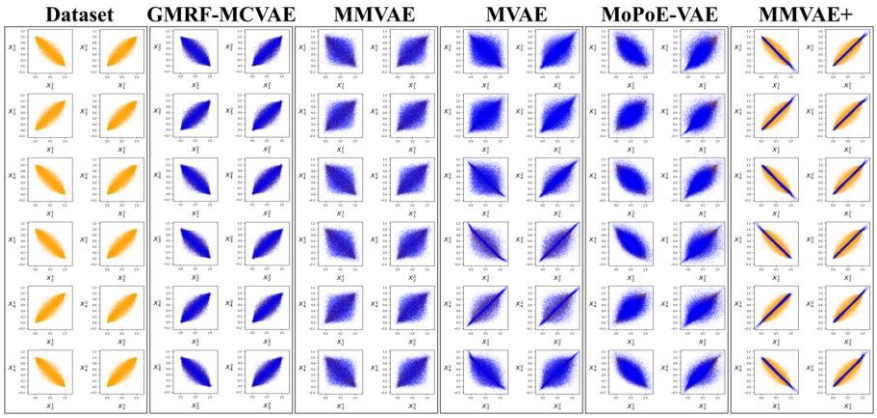


Figure 2 : Résultats qualitatifs des générations inconditionnelles sur le dataset de copules. Les distributions réelles sont représentées en orange et les distributions générées en bleu.

Model	Uncond. Gen.			Cond. Gen.		
	Dim 1	Dim 2	Mean	Dim 1	Dim 2	Mean
MVAE	2.7	3.2	2.9	3.0	3.1	3.1
MMVAE	5.2	4.5	4.8	5.4	4.7	5.0
MoPoE-VAE	1.9	2.6	2.2	6.0	5.6	5.9
MMVAE+	8.1	4.9	6.5	5.2	4.9	5.1
GMRF MCVAE	**0.7**	**0.95**	**0.86**	**2.6**	**2.7**	**2.6**

Table 1 : Résultats comparatifs sur le jeu de données basé sur les copules pour la génération inconditionnelle (Uncond.) et conditionnelle (Cond.) en utilisant des distances de Wasserstein mises à l'échelle (multipliées par 1000).

Sur le dataset PolyMNIST, nous observons que notre modèle produit des chiffres à la fois cohérents et diversifiés (Figure 3), tout en affichant des scores de cohérence et de similarité structurelle parmi les meilleurs (Table 2). Malgré des FID légèrement plus élevés que le

MMVAE+ et le MVAE, le GMRF-MCVAE préserve mieux les relations inter-composants.

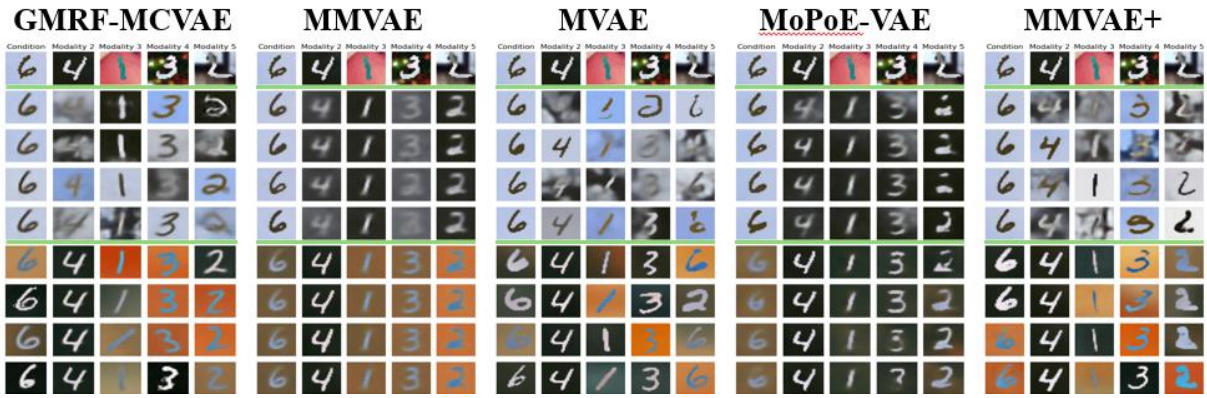


Figure 3 : Résultats de la génération conditionnelle pour PolyMNIST. La première ligne correspond à la condition, suivie des générations conditionnelles.

Model	Unconditional		Conditional		
	FID	Coherence	FID	Coherence	SSIM
MVAE	95.14	0.139	94.71	0.448	0.993
MMVAE	170.87	0.175	198.80	0.517	0.995
MoPoE-VAE	106.12	0.018	162.74	0.475	0.995
MMVAE+	**87.23**	0.210	**82.05**	0.856	0.994
GMRF MCVAE	118.21	**0.321**	180.76	**0.869**	0.995

Table 2 : Résultats sur PolyMNIST montrant le FID, la cohérence et le SSIM pour les différents modèles testés.

4 Système industriel complexe

Au-delà des résultats obtenus sur les datasets étudiés, l'objectif principal de ce travail réside dans l'application du GMRF-MCVAE à des systèmes industriels complexes. À cette fin, un dataset industriel basé sur l'épure de pneu, composé de six composants liés par des contraintes de conception (cf. Figure 4), est en cours d'étude pour valider le modèle dans un contexte réel.

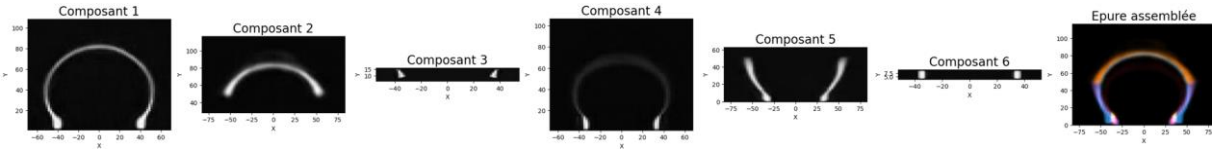


Figure 4 : Exemple de génération du GMRF-MCVAE sur le dataset d'épure, montrant les six composants générés et leur superposition.

5 Conclusion

Dans ce travail, nous proposons une architecture innovante intégrant des champs de Markov aléatoires gaussiens dans le cadre des Autoencodeurs Variationnels. Cette approche se révèle particulièrement adaptée pour capturer les relations complexes entre composants. Le modèle GMRF-MCVAE démontre des performances compétitives sur le benchmark PolyMNIST et surpasse les autres modèles multi-composants de l'état de l'art sur le dataset de copules, soulignant sa capacité à modéliser efficacement les dépendances inter-composants. En perspective, nous explorons actuellement l'application des modèles présentés dans ce travail ainsi que des architectures basées sur les modèles de diffusion sur le jeu de données d'épures de pneus. Par ailleurs, des travaux sur les métriques d'évaluation et sur l'interprétabilité des modèles sont également en cours d'étude.

Bibliographie

Rahimi, A., Khalil, A., Faisal, A., and Lai, K. W. (2022), CT-MRI dual information registration for the diagnosis of liver cancer: A pilot study using point-based registration, *Current Medical Imaging*, 18(1), pp. 61–66.

Puhr-Westerheide, D., Cyran, C. C., Sargsyan-Bergmann, J., Todica, A., Gildehaus, F.-J., Kunz, W. G., Stahl, R., Spitzweg, C., Ricke, J., and Kazmierczak, P. M. (2019), The added diagnostic value of complementary gadoxetic acid-enhanced MRI to 18 F-DOPA-PET/CT for liver staging in medullary thyroid carcinoma, *Cancer Imaging*, 19, pp. 1–10.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. (2023), Imagebind: One embedding space to bind them all, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190.

Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., and Huang, J. (2024), PIXIU: A Comprehensive Benchmark, Instruction Dataset and Large Language Model for Finance, *Advances in Neural Information Processing Systems*, 36.

Wu, M., and Goodman, N. (2018), Multimodal generative models for scalable weakly-supervised learning, *Advances in Neural Information Processing Systems*, 31.

Shi, Y., Paige, B., Torr, P., et al. (2019), Variational mixture-of-experts autoencoders for multi-modal deep generative models, *Advances in Neural Information Processing Systems*, 32.

Sutter, T. M., Daunhawer, I., and Vogt, J. E. (2021), Generalized multimodal ELBO, *arXiv preprint arXiv:2105.02470*.

Palumbo, E., Daunhawer, I., and Vogt, J. E. (2023), MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises, *The Eleventh International Conference on Learning Representations*.

Koller, D., and Friedman, N. (2009), *Probabilistic graphical models: Principles and techniques*, MIT Press.

Kingma, D. P., Welling, M., et al. (2019), An introduction to variational autoencoders, *Foundations and Trends in Machine Learning*, 12(4), pp. 307–392.

MacKenzie, D., and Spears, T. (2014), 'The formula that killed Wall Street': The Gaussian copula and modelling practices in investment banking, *Social Studies of Science*, 44(3), pp. 393–417.

Tedesco, P., Lenkoski, A., Bloomfield, H. C., and Sillmann, J. (2023), Gaussian copula modeling of extreme cold and weak-wind events over Europe conditioned on winter weather regimes, *Environmental Research Letters*, 18(3), p. 034008.

Daunhawer, I., Sutter, T. M., Chin-Cheong, K., Palumbo, E., and Vogt, J. E. (2022), On the limitations of multimodal VAEs.

Cobb, A., Roy, A., Elenius, D., Heim, F., Swenson, B., Whittington, S., Walker, J., Bapty, T., Hite, J., Ramani, K., et al. (2023). Aircraftverse: a large-scale multimodal dataset of aerial vehicle designs. Advances in Neural Information Processing Systems, 36, 44524–44543.

Analyse Théorique de la Convergence des Autoencodeurs Variationnels

Sobihan Surendran [1,2], Antoine Godichon-Baggioni [1], Sylvain Le Corff [1]

[1] *LPSM, Sorbonne Université, France*
[2] *LOPF, Califrais, Machine Learning Lab, Paris, France*
{*sobihan.surendran, antoine.godichon_baggioni, sylvain.le_corff*}*@sorbonne-universite.fr*

Résumé. Les Autoencodeurs Variationnels (VAE) sont des modèles génératifs largement utilisés pour l'échantillonnage à partir de distributions complexes. Bien qu'ils aient démontré une efficacité empirique dans de nombreuses applications, leur analyse théorique reste incomplète, en particulier concernant les garanties de convergence. Cet article apporte de nouveaux résultats en établissant des garanties de convergence non asymptotiques pour les VAE entraînés via la Descente de Gradient Stochastique et Adam. Nous établissons une vitesse de convergence de $\mathcal{O}(\log n/\sqrt{n})$, où n est le nombre d'itérations de l'algorithme d'optimisation, avec des dépendances explicites vis-à-vis de la taille des mini-batchs, du nombre d'échantillons variationnels et d'autres hyperparamètres. Notre analyse théorique s'applique aussi bien aux VAE linéaires qu'aux VAE gaussiens profonds, ainsi qu'à plusieurs variantes de VAE, dont β-VAE et IWAE. De plus, nous illustrons empiriquement l'impact des hyperparamètres sur la convergence, offrant de nouvelles perspectives pour la compréhension théorique de l'entraînement des VAE.

Mots-clés. Autoencodeurs Variationnels, Inférence Variationnelle, Optimisation Stochastique

Abstract. Variational Autoencoders (VAE) are popular generative models used to sample from complex data distributions. Despite their empirical success in various machine learning tasks, significant gaps remain in understanding their theoretical properties, particularly regarding convergence guarantees. This paper aims to bridge that gap by providing non-asymptotic convergence guarantees for VAE trained using both Stochastic Gradient Descent and Adam algorithms. We derive a convergence rate of $\mathcal{O}(\log n/\sqrt{n})$, where n is the number of iterations of the optimization algorithm, with explicit dependencies on the batch size, the number of variational samples, and other key hyperparameters. Our theoretical analysis applies to both Linear VAE and Deep Gaussian VAE, as well as several VAE variants, including β-VAE and IWAE. Additionally, we empirically illustrate the impact of hyperparameters on convergence, offering new insights into the theoretical understanding of VAE training.

Keywords. Variational Autoenconders, Variational Inference, Stochastic Optimization

1 Introduction

Probabilistic inference in generative models is a long-standing problem in particular for complex and high dimensional distributions. Many solutions to estimate the likelihood of data or to infer parameters are based on Importance Sampling, Sequential Monte Carlo (SMC) or Markov Chain Monte Carlo (MCMC)-based algorithms. Although their convergence properties have been analyzed in numerous research works, they face convergence problems and lead to slow training procedures in high-dimensional settings. Recently, many advances have been proposed in generative modeling, leading to the emergence of novel approaches such as Variational Autoencoders (VAE) [Kingma and Welling, 2014, Rezende et al., 2014, Ranganath et al., 2014], Generative Adversarial Networks (GAN) [Goodfellow et al., 2014], and Probabilistic and Score-Based Diffusion Models [Ho et al., 2020, Song et al., 2021]. These models offer alternative training paradigms that address many of the limitations associated with MCMC-based approaches. Among these, VAE are notable for their ability to learn low dimensional latent variables, while maintaining a clear probabilistic interpretation, enabling both scalable training and meaningful generative modeling.

VAE have been successfully applied in many different contexts such as text generation, image generation, image segmentation, representation learning, music generation, dimensionality reduction and anomaly detection. There are several popular variants of VAE, including IWAE [Burda et al., 2016], β-VAE [Higgins et al., 2017], VQ-VAE [Van Den Oord et al., 2017], and Conditional VAE [Sohn et al., 2015].

Theoretical properties of Variational Inference have only recently been analyzed. For instance, Chérief-Abdellatif et al. [2022] provides generalization bounds on the theoretical reconstruction error using PAC-Bayes theory, while Huggins et al. [2020] offers variational error bounds for posterior mean and uncertainty estimates. Mbacke et al. [2024] extend these analyses by offering statistical guarantees for reconstruction, generation, and regeneration. Furthermore, Tang and Yang [2021] highlights the importance of covariance matrices in Gaussian encoders and decoders, providing variational excess risk bounds for Empirical Bayes Variational Autoencoders (EBVAE) applied to Gaussian models. The authors provide an oracle result for the EBVAE estimator which is crucial to prove the consistency of the estimator. Additionally, theoretical insights into posterior collapse in VAE are provided by [Razavi et al., 2019, Lucas et al., 2019, Wang and Ziyin, 2022], particularly in the context of Linear VAE. More recently, convergence results for Black-Box Variational Inference have been established by [Domke et al., 2023, Kim et al., 2024].

While many theoretical guarantees have been developed for Variational Inference, obtaining convergence guarantees and non-asymptotic convergence rates for VAEs remains mainly an open problem. In this paper, we address this gap by providing non-asymptotic convergence guarantees for VAE, when using standard Stochastic Gradient Descent (SGD) [Robbins and Monro, 1951] and Adam [Kingma and Ba, 2015] optimization algorithms. Our results also yield new insights into the theoretical properties of Variational Inference.

2 Convergence Analysis of VAE in the General Setting

Let $\mathsf{X} \subseteq \mathbb{R}^{d_x}$ and $\mathsf{Z} \subseteq \mathbb{R}^{d_z}$ denote the data space and the latent space, respectively. We first define the expected Evidence Lower Bound (ELBO) over the data distribution π as follows:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{\pi,\phi}\left[\log \frac{p_\theta(X, Z)}{q_\phi(Z|X)}\right] = \mathbb{E}_\pi\left[\mathcal{L}(\theta, \phi; X)\right] \ .$$

In order to optimize θ and ϕ, one needs to compute the gradients of the ELBO with respect to these parameters. Under classical regularity assumptions, the gradient with respect to θ is given by:

$$\nabla_\theta \mathcal{L}(\theta, \phi) = \mathbb{E}_{\pi,\phi}\left[\nabla_\theta \log p_\theta(X, Z)\right] \ .$$

Computing the gradient with respect to the variational parameters ϕ is more challenging since the inner expectation depends on q_ϕ. The reparametrization trick addresses this issue by expressing the latent variable z as a deterministic transform $z = g(\varepsilon, \phi)$, where ε is an auxiliary independent random variable drawn from a known distribution p_ε. Using this trick, the pathwise gradient [Kingma and Welling, 2014, Rezende et al., 2014] of the ELBO can be expressed as:

$$\nabla_\phi \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{p_\varepsilon}\left[\nabla_z \log w_{\theta,\phi}(x, z) \nabla_\phi g(\varepsilon, \phi)\right] - \mathbb{E}_{p_\varepsilon}\left[\nabla_\phi \log q_\phi(g(\varepsilon, \phi)|x)\right] \ ,$$

where $w_{\theta,\phi}(x, z) = p_\theta(x, z)/q_\phi(z|x)$ the unnormalized importance weights and $\mathbb{E}_{p_\varepsilon}$ is the expectation under the law of ε when $\varepsilon \sim p_\varepsilon$.

SGD is a widely used method for training statistical models based on deep architectures. It produces a sequence of parameter estimates as follows: $(\theta_0, \phi_0) \in \Theta \times \Phi$ and for all $k \in \mathbb{N}$,

$$(\theta_{k+1}, \phi_{k+1}) = (\theta_k, \phi_k) + \gamma_{k+1} \widehat{\nabla}_{\theta,\phi} \mathcal{L}(\theta_k, \phi_k; \mathcal{D}_{k+1}) \ , \tag{1}$$

where $\widehat{\nabla}_{\theta,\phi} \mathcal{L}(\theta_k, \phi_k; \mathcal{D}_{k+1})$ denotes an estimator of the gradient, $\mathcal{D}_{k+1}$ is the mini-batch of data used at iteration $k + 1$ and for all $k \geq 1$, $\gamma_k > 0$ is the learning rate.

A crucial step in establishing the convergence rate of SGD is the smoothness of the expected ELBO. Building on this property, we derive the following result, which provides a convergence rate of $\mathcal{O}(\log n/\sqrt{n})$, where n is the number of iterations of the optimization algorithm, with explicit dependency on the batch size B and the number of variational samples K used at each gradient step.

Theorem 2.1. *Let $(\theta_n, \phi_n) \in \Theta \times \Phi$ be the n-th iterate of the recursion (1), where $\gamma_n = C_\gamma n^{-1/2}$ with $C_\gamma > 0$. Assume that $\sigma^2 = \sup_{\theta\in\Theta,\phi\in\Phi} \mathbb{E}_\pi[\mathbb{E}_{q_\phi(\cdot|X_i)}[\|\tilde{g}_{i,\ell} - \nabla_{\theta,\phi}\mathcal{L}(\theta, \phi)\|^2]] < +\infty$. For all $n \geq 1$, let $R \in \{0, \ldots, n\}$ be a uniformly distributed random variable. Then, under standard regularity assumptions and assuming $C_\gamma \leq 1/L$, we have:*

$$\mathbb{E}\left[\|\nabla_{\theta,\phi}\mathcal{L}\left(\theta_R, \phi_R\right)\|^2\right] \leq \frac{2\mathcal{L}^* + LC_\gamma\sigma^2 \log n/(BK)}{C_\gamma\sqrt{n}} \ ,$$

where $\mathcal{L}^* = \mathcal{L}\left(\theta^*, \phi^*\right) - \mathcal{L}\left(\theta_0, \phi_0\right)$ and L is the smoothness constant of $\mathcal{L}$.

We establish that our results apply to both Linear VAE and Deep Gaussian VAE, verifying that the assumptions underlying our main theorems hold for standard neural network architectures and discussing the choice of hyperparameters. The deep Gaussian VAE consists of a decoder and an encoder such that $p_\theta(x|z) = \mathcal{N}(x; G_\theta(z), c^2 \mathrm{I}_{d_x})$ and $q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \Sigma_\phi(x))$. To support our assumptions, we introduce a generalized soft-clipping activation function that is both Lipschitz continuous and smooth. For $s_1, s_2 \in \mathbb{R}$ with $s_1 \leq s_2$ and $s \in \mathbb{R}_+^*$, the generalized soft-clipping activation function is defined as:

$$f(x) = \frac{1}{s} \log \left(\frac{1 + \mathrm{e}^{s(x-s_1)}}{1 + \mathrm{e}^{s(x-s_2)}} \right) + s_1.$$

Under these conditions, we derive a convergence rate of $\mathcal{O}\left(N a^{2(N-1)} \log n / \sqrt{n}\right)$ for deep Gaussian VAE, where N denotes the total number of layers in the encoder and decoder, assuming the existence of a constant a such that $\|\theta\|_\infty + \|\phi\|_\infty \leq a$ for all $\theta \in \Theta$ and $\phi \in \Phi$.

3 Some Variants of VAE

We derive convergence rates for popular VAE variants, including β-VAE and IWAE.

β-VAE. β-VAE [Higgins et al., 2017] is a variant of VAE introducing a parameter β to control the trade-off between the reconstruction term and the regularization of the latent space. The ELBO for β-VAE is given by:

$$\mathcal{L}_\beta(\theta, \phi; x) = \mathbb{E}_{q_\phi(\cdot|x)} \left[\log p_\theta(x|Z) \right] - \beta D_{\mathrm{KL}}(q_\phi(\cdot|x)||p) \ ,$$

where the Lagrangian multiplier β is considered as a hyperparameter. We observe that if $\beta < \infty$, β-VAE converges to a critical point of the expected ELBO at a rate of $\mathcal{O}\left(\log n / \sqrt{n}\right)$. Selecting a sufficiently small β leads to faster convergence and can help prevent posterior collapse [Wang and Ziyin, 2022].

IWAE. The Importance Weighted Autoencoder (IWAE) [Burda et al., 2016] is another extension of the VAE that incorporates importance weighting to obtain a tighter ELBO. The IWAE objective function is defined as:

$$\mathcal{L}_K^{\mathsf{IWAE}}(\theta, \phi) = \mathbb{E}_\pi \left[\mathbb{E}_{q_\phi^{\otimes K}(\cdot|X)} \left[\log \frac{1}{K} \sum_{\ell=1}^{K} \frac{p_\theta(X, Z^{(\ell)})}{q_\phi(Z^{(\ell)}|X)} \right] \right] \ ,$$

where K corresponds to the number of samples drawn from the variational posterior distribution. We achieve a convergence rate similar to that of VAE and β-VAE, with IWAE converging to a critical point of the expected ELBO at a rate of $\mathcal{O}\left(K^{-1} \log n / \sqrt{n}\right)$. Notably, as K increases, the convergence rate improves and becomes nearly inversely proportional to K.

4 Experiments

We illustrate our convergence results by analyzing empirically the impact of hyperparameters, particularly the regularization parameter β in β-VAE, the number of samples from the

variational distribution in IWAE. We conduct our experiments on the CelebA dataset and use a Convolutional Neural Network (CNN) architecture with the generalized soft-clipping activation functions for both the encoder and decoder networks. The latent space dimension is set to 100. We estimate the log-likelihood using the VAE, β-VAE, and IWAE models, all of which are trained for 100 epochs. Training is performed with Adam optimizer and learning rate decay defined as $\gamma_n = C_\gamma/\sqrt{n}$, where $C_\gamma = 0.001$. The momentum parameters are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the regularization parameter δ is fixed at 10^{-8}. Note that while all figures are plotted with respect to epochs, n here denotes the number of gradient updates.

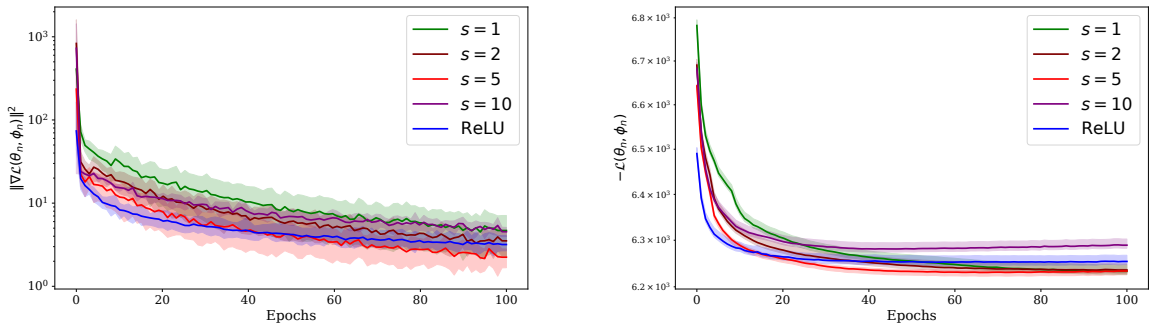


Figure 1: Squared norm of gradients and Negative ELBO on the test set of the CelebA for VAE trained with Adam and generalized soft-clipping activation function. Bold lines represent the mean over 5 independent runs.

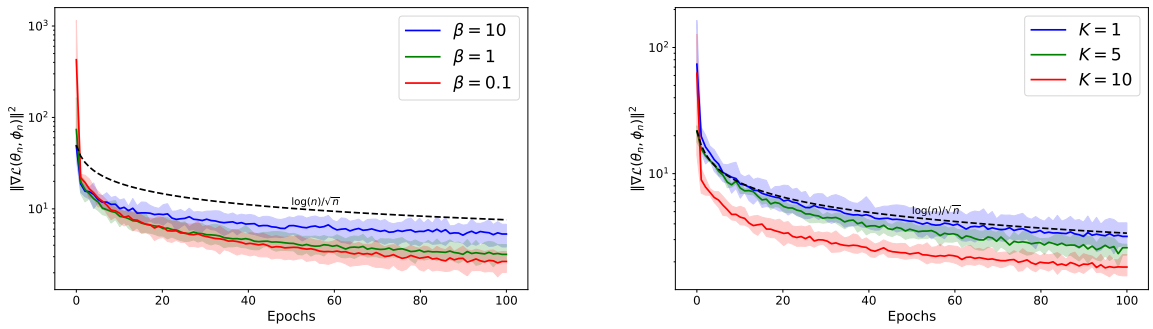


Figure 2: $\|\nabla\mathcal{L}(\theta_n, \phi_n)\|^2$ in β-VAE (on the left) and IWAE (on the right) trained with Adam. Bold lines represent the mean over 5 independent runs. The dashed curves correspond to the expected convergence rate $\mathcal{O}(\log n/\sqrt{n})$.

For the first experiment, we illustrate our convergence results of the standard VAE with our choice of activation functions. Figure 1 shows the squared norm of the gradients $\|\nabla\mathcal{L}(\theta_n, \phi_n)\|^2$ and the Negative ELBO on the test dataset for both ReLU and the generalized soft-clipping activation function with various values of s. We observe a similar convergence rate for all values of s. However, selecting an appropriate value of s is crucial for achieving optimal convergence. Theoretically, smaller values of s should result in slower convergence

rates, but in practice, choosing a very large s can introduce numerical instabilities due to the exponential term in the generalized soft-clipping function. Then, $s = 5$ appears to be a reasonable choice, balancing convergence rate and numerical stability. Moreover, our choice of activation functions leads to improved convergence rates compared to the standard VAE with ReLU. While our theoretical analysis does not directly apply to ReLU, experimental results indicate that similar convergence performance can be achieved.

Figure 2 illustrates the squared norm of the gradients $\|\nabla\mathcal{L}(\theta_n, \phi_n)\|^2$ for both the β-VAE and IWAE models, evaluated across different values of β and K. The standard VAE is a special case of these models, corresponding to $\beta = 1$ in the β-VAE and $K = 1$ in the IWAE. As expected, we observe that for the β-VAE, smaller values of β lead to faster convergence. Similarly, for the IWAE, increasing the value of K results in faster convergence.

References

Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

B.-E. Chérief-Abdellatif, Y. Shi, A. Doucet, and B. Guedj. On PAC-Bayesian reconstruction guarantees for VAEs. In *International Conference on Artificial Intelligence and Statistics*, pages 3066–3079. PMLR, 2022.

J. Domke, R. Gower, and G. Garrigos. Provable convergence guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

J. Huggins, M. Kasprzak, T. Campbell, and T. Broderick. Validated variational inference via practical posterior error bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 1792–1802. PMLR, 2020.

K. Kim, Y. Ma, and J. Gardner. Linear convergence of black-box variational inference: Should we stick the landing? In *International Conference on Artificial Intelligence and Statistics*, pages 235–243. PMLR, 2024.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

J. Lucas, G. Tucker, R. B. Grosse, and M. Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

S. D. Mbacke, F. Clerc, and P. Germain. Statistical guarantees for variational autoencoders using pac-bayesian theory. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *International Conference on Artificial Intelligence and Statistics*, pages 814–822. PMLR, 2014.

A. Razavi, A. v. d. Oord, B. Poole, and O. Vinyals. Preventing posterior collapse with delta-vaes. In *International Conference on Learning Representations*, 2019.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286. PMLR, 2014.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

R. Tang and Y. Yang. On empirical bayes variational autoencoder: An excess risk bound. In *Conference on Learning Theory*, pages 4068–4125. PMLR, 2021.

A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Z. Wang and L. Ziyin. Posterior collapse of a linear latent variable model. In *Advances in Neural Information Processing Systems*, volume 35, pages 37537–37548, 2022.

IA pour la découverte scientifique : prospective trans sociétés savantes

AI as a Scientific Pilot?

Karteek Alahari [1] & Christophe Biernacki [2]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*
[2] *Inria, CNRS, Laboratoire de Mathématiques Painlevé, U. Lille, F59650 V. d'Ascq, France*

Recent developments in the field of artificial intelligence (AI) have produced phenomenal tools. These range from tools to find and summarize relevant research papers [1, 2], applications to find potential trends [6] or novel research hypotheses or proposals [7, 9], assistive tools for writing [3], tools to measure compliance of open science requirements [4], and even autonomous systems that generate research papers [10]. In fact, one such AI-generated paper has been accepted at an ICLR workshop earlier this year. This was done as an experiment in controlled settings in collaboration with the workshop organizers [5]. There are also efforts to evaluate the ability of AI systems to replicate research results [8]. Given all these advances, we pose the question of whether AI systems can take the role of a principal investigator (PI) and pilot scientific research. In this talk, we will summarize relevant work in this context, with a focus on open questions to the scientific community.

Keywords. Generative AI, AI scientist.

Les avancées récentes dans le domaine de l'intelligence artificielle (IA) ont conduit à l'émergence d'outils remarquables aux applications variées. Ces outils incluent notamment des systèmes permettant d'identifier et de synthétiser des publications scientifiques pertinentes [1, 2], des applications visant à détecter des tendances émergentes [6] ou à formuler de nouvelles hypothèses ou propositions de recherche [7, 9], des dispositifs d'assistance à la rédaction scientifique [3], des outils d'évaluation de la conformité aux principes de la science ouverte [4], ainsi que des systèmes autonomes capables de générer des articles scientifiques de manière indépendante [10]. À titre d'illustration, un article généré par une IA a récemment été accepté dans le cadre d'un atelier de la conférence ICLR, dans un contexte expérimental encadré et en collaboration avec les organisateurs de l'événement [5]. Par ailleurs, des travaux sont en cours afin d'évaluer la capacité des systèmes d'IA à reproduire les résultats expérimentaux publiés [8]. Face à ces développements, une question fondamentale émerge : les systèmes d'intelligence artificielle peuvent-ils assumer le rôle de chercheur principal (PI) et conduire de manière autonome un programme de recherche scientifique ? Cette présentation propose une synthèse des travaux existants sur cette problématique, en mettant en avant les questions ouvertes et les enjeux épistémologiques associés, à l'attention de la communauté scientifique.

Mots-clés. IA Générative, IA scientifique.

References

[1] https://www.paperpilot.xyz.

[2] https://www.aimodels.fyi.

[3] https://paperpal.com.

[4] https://dataseer.ai.

[5] https://github.com/SakanaAI/AI-Scientist-ICLR2025-Workshop-Experiment.

[6] NotebookLM. https://notebooklm.google/.

[7] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an AI co-scientist. https://arxiv.org/abs/2502.18864, 2025.

[8] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. Paperbench: Evaluating AI's ability to replicate AI research. https://arxiv.org/abs/2504.01848, 2025.

[9] Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. Many heads are better than one: Improved scientific idea generation by a llm-based multi-agent system. https://arxiv.org/abs/2410.09403v2, 2024.

[10] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. Sakana AI Blog Post. Available online: https://pub.sakana.ai/ai-scientist-v2/paper, April 2025.

Title : AI4Science: From Equations to Algorithms

Patrick Gallinari, Criteo AI Lab, Paris and Sorbonne Universite

patrick.gallinari@sorbonne-universite.fr

https://pages.isir.upmc.fr/gallinari/

https://dblp.org/pid/g/PatrickGallinari.html

Abstract:

For centuries, physics has advanced through the explicit formulation of laws—equations derived from first principles. Today, however, we are witnessing a profound shift: the rise of data-driven models that learn physical behavior directly from observations, bypassing the need for closed-form equations. This talk explores the emerging paradigm of **AI4Science**, where neural networks and machine learning algorithms increasingly serve as surrogate models, predictive engines, and even tools for scientific discovery. We will examine the implications of this transition—from interpretability and generalization to the nature of scientific understanding itself—and discuss how this new era challenges traditional roles in science.

Keywords : AI4Science ; Deep learning ; Machine Learning for Physics

Presentation

This presentation explores the profound impact of machine learning on scientific discovery, marking a shift from AI as a computational aid to a fundamental scientific paradigm. We will delve into how deep learning, in particular, is reshaping the very foundations of scientific modeling, especially for intricate physical systems, by enabling the development of powerful empirical models capable of learning complex patterns and potentially generalizing across diverse scenarios

To illustrate this transformative shift, the first part of the presentation will survey compelling examples from a range of scientific domains, including climate and weather forecasting, materials science, biology and drug discovery, and physics.

The second part of the presentation will describe key research topics developed in our team over the last few years, focusing on the modeling of physical dynamical systems. We will highlight the deployment and challenges of machine learning in this domain, with a particular emphasis on three main areas:

- Integrating Physics and ML (Hybrid Systems): Exploring the integration of physical knowledge and machine learning through hybrid systems, illustrating how neural components can enhance and complement traditional physics-based models.
- Domain Generalization Challenges: Addressing the critical challenge of domain generalization in modeling complex physical processes: developing neural models that can reliably perform across varying environments and generalize beyond their training data distribution.
- Neural Operators for Function Space Mapping: Presenting the development of Neural Operators, which learn mappings between function spaces, offering a powerful approach for modeling physical systems independent of specific geometries or resolutions.

LES AGENTS CONVERSATIONNELS DANS LA DÉMARCHE SCIENTIFIQUE : POTENTIEL ET LIMITES

Adrien Guille

Université Lumière Lyon 2, UR ERIC 3983
France, adrien.guille@univ-lyon2.fr

Résumé. Les cas d'usage potentiels des agents conversationnels dans le cadre de la démarche scientifique semblent nombreux, par exemple : formulation automatique d'hypothèses, analyse automatique des résultats expérimentaux voire même, dans certains domaines, programmation automatique des expériences, rédaction d'articles assistée ou automatisée. Après avoir rappelé brièvement les fondements des agents conversationnels et des grands modèles de langue sur lesquels ils reposent, cette présentation discutera de leurs limites et des précautions à prendre dans un contexte scientifique. En effet, ces modèles, sans représentation logique ou causale des phénomènes, ne « comprennent » pas véritablement les concepts et raisonnements scientifiques et sont sujets aux hallucinations, *i.e.* la génération de réponses en apparence plausibles mais en réalité incorrectes. De plus, le processus par lequel ils génèrent leurs réponses ne permet pas aisément de retracer les sources mobilisées. Par ailleurs, les données utilisées pour les différentes phases d'estimation des paramètres de ces modèles peuvent induire des biais dans les réponses.

Mots-clés. Agent conversationnel, grand modèle de langue, cas d'usage, limites

Abstract. The potential use cases for conversational agents in the context of the scientific process appear numerous, for example: automatic hypothesis generation, automated analysis of experimental results, and even, in certain fields, the automatic coding of experiments, assisted or automated article writing. After briefly reviewing the foundations of chatbots and the large language models they rely on, this presentation will discuss their limitations and the precautions that should be taken in a scientific context. Indeed, these models, lacking a logical or causal representation, do not truly "understand" scientific concepts and reasoning, and are prone to hallucinations — *i.e.* generating responses that seem plausible but are actually incorrect. Furthermore, the process by which they generate their responses does not easily allow for tracing the sources used. Additionally, the data used during the various parameter estimation phases of these models may introduce biases into the responses.

Keywords. Chatbot, large language model, use cases, shortcomings

IA responsable

Fundamental Limits of Membership Inference Attacks on Machine Learning Models

Eric Aubinais [1] & Elisabeth Gassiat [1] & Pablo Piantanida [2]

[1] *Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, France,* {*prenom.nom*}*@universite-paris-saclay.fr*
[2] *ILLS - International Laboratory on Learning Systems, MILA - Quebec AI Institute, Montreal (QC), Canada, CNRS, CentraleSupélec Université Paris-Saclay, pablo.piantanida@mila.quebec*

Résumé. Les attaques par inférence d'appartenance (MIA) peuvent révéler si un point de données particulier faisait partie de l'ensemble de données d'apprentissage, exposant potentiellement des informations sensibles sur les individus. Cet article fournit des garanties théoriques en explorant les limites statistiques fondamentales associées aux attaques par inférence sur les modèles de machine learning en général. Plus précisément, nous commençons par dériver la quantité statistique qui régit l'efficacité et le succès de ces attaques. Nous prouvons ensuite théoriquement que dans un cadre de régression (non-)linéaire avec des algorithmes présentant de l'overfitting, certaines MIAs peuvent avoir une forte probabilité de réussite. Enfin, nous étudions plusieurs situations pour lesquelles nous fournissons un contrôle cette quantité d'intérêt. Il est intéressant de noter que nos résultats indiquent que la discrétisation des données pourrait améliorer la sécurité de l'algorithme. En particulier, il est démontré qu'il est limité par une constante qui quantifie la diversité de la distribution des données sous-jacentes.

Mots-clés. Attaques par inférence d'appartenance, Limitations Statistiques, Sécurité des données, Bornes théoriques

Abstract. Membership inference attacks (MIA) can reveal whether a particular data point was part of the training dataset, potentially exposing sensitive information about individuals. This article provides theoretical guarantees by exploring the fundamental statistical limitations associated with MIAs on machine learning models at large. More precisely, we first derive the statistical quantity that governs the effectiveness and success of such attacks. We then theoretically prove that in a (non-)linear regression setting with overfitting algorithms, attacks may have a high probability of success. Finally, we investigate several situations for which we provide bounds on this quantity of interest. Interestingly, our findings indicate that discretizing the data might enhance the algorithm's security. Specifically, it is demonstrated to be limited by a constant, which quantifies the diversity of the underlying data distribution.

Keywords. Membership Inference Attacks, Statistical Limitations, Privacy, Theoretical Performance Bounds, Overfitting, Trustworthy Machine Learning.

1 Problem Setup

Let $\mathbf{z} \coloneqq (z_1, \cdots, z_n) \in \mathcal{Z}^n$ be a dataset of independent and identically distributed (*i.i.d.*) random variables from a distribution P. Let $\mathcal{F} \coloneqq \{\mu_\theta : \theta \in \Theta\}$ be a set of predictors where Θ is the space of parameters. For instance, $\mathcal{F}$ can be the set of all neural networks with fixed architecture. The predictor is then identified to its parameters $\hat{\theta}_n \in \Theta$ learned from $\mathbf{z}$ through a learning procedure $\mathcal{A} : \bigcup_{k>0} \mathcal{Z}^k \to \mathcal{P}' \subseteq \mathcal{P}(\Theta)$, that is $\hat{\theta}_n$ follows the distribution $\mathcal{A}(\mathbf{z})$ conditionally to $\mathbf{z}$, which we assume we have access to. Here, $\mathcal{P}(\Theta)$ is the set of all distributions on Θ, and $\mathcal{P}'$ is the range of $\mathcal{A}$. We will further assume that the learning procedure $\mathcal{A}$ can be expressed as a function of the empirical distribution of the training dataset. We therefore focus on MIAs, the ability of recovering membership to the dataset $\mathbf{z}$ of a test point $\tilde{z} \in \mathcal{Z}$ from the predictor $\mu_{\hat{\theta}_n}$. Within this framework, we consider MIAs as functions of the parameters and the test point whose outputs are 0 or 1.

Definition 1 (**Membership Inference Attack - MIA**) *Any measurable map $\phi : \Theta \times \mathcal{Z} \to \{0, 1\}$ is called a **Membership Inference Attack**.*

Importantly, we assume that MIAs might access more information, including randomization, the learning procedure $\mathcal{A}$ and/or the distribution of the data P.
We suppose that the test point $\tilde{z}$ is to be drawn from P independently from the samples $\{z_1, \ldots, z_n\}$ with probability $\nu \in (0, 1)$. Otherwise, conditionally to $\mathbf{z}$, we set $\tilde{z}$ to any z_j each with uniform probability $1/n$. Letting U be a random variable with distribution $\hat{P}_n \coloneqq \frac{1}{n}\sum_{j=1}^n \delta_{z_j}$ conditionally to $\mathbf{z}$, z_0 to be drawn independently from P and T be a random variable having Bernoulli distribution with parameter ν and independent of any other random variables, we shall set $\tilde{z} \coloneqq Tz_0 + (1-T)U$. The probability ν represent the **fraction of non-member** samples when evaluating MIAs, which is usually unknown in practice,

A sensible choice to evaluate the performance of an MIA would be to consider its probability of successfully guessing the membership, i.e.

$$\begin{aligned}\mathbb{P}(\phi(\hat{\theta}_n, \tilde{z}) = 1 - T) &= \mathbb{P}(\phi(\hat{\theta}_n, z_0) = 0, T = 1) + \mathbb{P}(\phi(\hat{\theta}_n, U) = 1, T = 0) \\ &= \mathbb{E}\left[TNR + TPR\right],\end{aligned}$$

where TPR (resp. TNR) is the True Positive Rate (resp. True Negative Rate) of the MIA ϕ. However, in an MIA setting, the TPR is arguably more important than the TNR. Therefore, we define the **importance of the TPR** as a real number $\lambda > 0$, and we measure the performance of an MIA by its weighted probability of successfully guessing the membership of the test point.

Definition 2 (**Accuracy of an MIA**) *The **accuracy of an MIA** ϕ is defined as*

$$Acc_n(\phi; P, \mathcal{A}) \coloneqq \mathbb{P}\left(\phi(\hat{\theta}_n, z_0) = 0, T = 1\right) + \lambda \mathbb{P}\left(\phi(\hat{\theta}_n, U) = 1, T = 0\right), \tag{1}$$

where the probability is taken over all randomness.

The quantity λ represents the importance accorded to the TPR over the TNR in the prediction of an MIA. This is usually a meta-parameter which is user-specified.

Remark 1 The accuracy of an MIA scales from 0 to $\nu + \lambda(1-\nu)$. Constant MIAs $\phi_0 \equiv 0$ and $\phi_1 \equiv 1$ have respectively an accuracy equal to ν and $\lambda(1-\nu)$, which means that we always can build an MIA with accuracy of at least $\max(\nu, \lambda(1-\nu))$ and any MIA performing worse than this quantity is irrelevant to use. Particularly, this means that we have

$$\max(\nu, \lambda(1-\nu)) \leq \sup_{\phi} \text{Acc}_{\nu,\lambda,n}(\phi; P, \mathcal{A}) \leq \nu + \lambda(1-\nu).$$

Therefore, we now define the **Membership Inference Security** of a learning procedure as a quantity summarizing the amount of security of the learning procedure against MIAs.

Definition 3 (**Membership Inference Security - MIS**) *The Membership Inference Security of a learning procedure $\mathcal{A}$ is*

$$Sec_{\nu,\lambda,n}(P, \mathcal{A}) \coloneqq \frac{1}{\min(\nu, \lambda(1-\nu))} \left(\nu + \lambda(1-\nu) - \sup_{\phi} Acc_{\nu,\lambda,n}(\phi; P, \mathcal{A}) \right), \tag{2}$$

where the sup is taken over all MIAs.

From Remark 1, we see that the MIS has been defined to ensure that it scales from 0 (the best MIA approaches perfect guess of membership) to 1 (MIAs can not do better than ϕ_0 and ϕ_1).

2 Results

Our objective is then to control the MIS in different scenarios. Note that the MIS has been introduced to encompass the MIA-wise security of a learning procedure when the TPR has an importance of λ over the TNR. Specifically, it has been defined to scale from 0 to 1, and therefore to be interpretable, i.e. 0 meaning no security and 1 full security. A general observation on the MIS, is that it is well-behaved in the following sense,

Theorem 1 *The quantity $1 - Sec_{\nu,\lambda,n}(P, \mathcal{A})$ is an $f-$divergence between $\mathbb{P}_{(\hat{\theta}_n, z_0)}$ and $\mathbb{P}_{(\hat{\theta}_n, z_1)}$ with as generator the function,*

$$f_{\lambda,\nu}(x) = (1/2) \max(1, \frac{\nu}{\lambda(1-\nu)}) \left[|x - \lambda(1-\nu)/\nu| - |1 - \lambda(1-\nu)/\nu| \right].$$

Theorem 1 states that $1 - \text{Sec}_{\nu,\lambda,n}(P, \mathcal{A})$ has the same properties as most information theory measures such as the KL divergence. In particular, an additional observation is that $1 - \text{Sec}_{\nu,\lambda,n}(P, \mathcal{A})$ will satisfy composition theorems as differentially private mechanisms do (Dwork (2014)).

2.1 Low Security Regime

Assume that $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{Y}$ and that the learning procedure $\mathcal{A}$ produces overfitting parameters $\hat{\theta}_n$. Assume furthermore that $\mathcal{A}$ minimizes $L_n : \theta \mapsto \frac{1}{n} \sum_{i=1}^{n} l_\theta(x_i, y_i)$ where $l_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$

is a loss function and $z \coloneqq (x, y)$. An empirically known fact is that overfitting algorithms tend to leak more knowledge.

Definition 4 ($(\varepsilon, 1-\alpha)-$**overfitting**) *We say that the algorithm $\mathcal{A}$ is $(\varepsilon, 1-\alpha)$-overfitting for some $\varepsilon \in \mathbb{R}^+$ and $\alpha \in (0,1)$ when*

$$\mathbb{P}\left(l_{\hat{\theta}_n}(x_1, y_1) \leq \varepsilon\right) \geq 1 - \alpha, \tag{3}$$

where the probability is taken over the data and the randomness of $\mathcal{A}$.

In particular, it can be shown that if a learning procedure stops as soon as $L_n \leq \eta$ for some $\eta > 0$, then it is $(\varepsilon, 1-\alpha)-$overfitting for any (ε, α) such that $\eta = \varepsilon\alpha$.

Theorem 2 (**Overfitting induces lack of security**) *Suppose that the following assertions holds,*

- *For all $\eta > 0$, the learning procedure $\mathcal{A}_\eta$ stops as soon as $L_n \leq \eta$.*
- *A version of the conditional distribution of y given x is absolutely continuous with respect to the Lebesgue measure.*
- *There exists some continuous function ω such that $\ell_\theta(x, y) = \omega(y, \Psi_\theta(x))$.*

Then it holds that

$$\lim_{\eta \to 0+} Sec_{\nu,\lambda,n}(P, \mathcal{A}) = 0. \tag{4}$$

Theorem 2 states that loss-based MIAs $\phi_\varepsilon(\theta, z) = 1_{\ell_\theta(x,y) \leq \varepsilon}$ achieve high accuracy. Additionally, if the first point does not hold but Definition 4 still hold, then it can be shown that the MIS of the learning procedure satisfies $\lim_{\varepsilon \to 0} \mathrm{Sec}_{\nu,\lambda,n}(P, \mathcal{A}) \leq \max(1, \lambda(1-\nu)/\nu)\alpha$.

2.2 High security regime

Empirical means. Alternatively, in simple models such as a linear regression or solving equations, a remarkable property of the learned parameters $\hat{\theta}_n$ is that they can be expressed as a function of an empirical mean of the data,

$$\hat{\theta}_n = F\left(\frac{1}{n}\sum_{j=1}^{n} L(z_j)\right),$$

for some functions $L : \mathcal{Z} \to \mathbb{R}^d$ and $F : \mathbb{R}^d \to \mathbb{R}^q$. In particular, this means that the learning procedure can be written as $\mathcal{A} : (z_1, \cdots, z_n) \mapsto \delta_{F\left(\frac{1}{n}\sum_{j=1}^{n} L(z_j)\right)}$, where δ stand for the Dirac mass at θ. In that context, we have the following result.

Theorem 3 *Suppose that the distribution of $L(z_1)$ has a non zero absolutely continuous part with respect to the Lebesgue measure, and a third finite moment. Then there exists a constant*

$c_{L,P} > 0$ *such that*

$$Sec_{\nu,\lambda,n}(P, \mathcal{A}) \geq 1 - \max(1, \nu/\lambda(1-\nu)) \left[c_{L,P} + \frac{\sqrt{d}}{2\sqrt{n}} \right] n^{-1/2}. \tag{5}$$

Discrete Data Supplementary, when the data is assumed to have a discrete distribution $P \coloneqq \sum_{k=1}^{K} p_k \delta_{u_k}$, the MIS behave similarly up to constants. As the theoretical value of ν can not be known in practice, we state the following result by taking the worst case scenario in terms of MIS. Defining $C_K(P) \coloneqq \sum_{k=1}^{K} \sqrt{p_k(1-p_k)}$, the following result holds.

Theorem 4 (**Discrete data**) *Assume that* $C_K(P) < \infty$, $n \geq 5$ *and* $n > 1/p_j$ *for all* $j = 1, \ldots, K$. *Then there exists universal constants* $c \geq 0.29$ *and* $c' \leq 0.44$ *such that*

$$\min_{\nu,\mathcal{A}} Sec_{\nu,\lambda,n}(P, \mathcal{A}) = 1 - \varepsilon_n,$$

where ε_n *satisfies*

$$cC_K(P)n^{-1/2} \leq \varepsilon_n \leq c'C_K(P)n^{-1/2}.$$

If $C(P) < \infty$ *but the condition on* n *does not hold, we still have* $\min_{\nu,\mathcal{A}} Sec_{\nu,\lambda,n}(P, \mathcal{A}) \geq 1 - \frac{C_K(P)}{2} n^{-1/2}$.

Theorems 3 and 4 state that when the learning procedure that can be formulated as a function of an empirical mean of the data or when the data is discrete, there exists a constant C such that $\mathrm{Sec}_{\nu,\lambda,n}(P, \mathcal{A}) = 1 - Cn^{-1/2}$ for the least secure algorithm in the worst theoretical value of ν. In particular, this means that for any learning procedure $\mathcal{A}$, no matter the theoretical value of ν, the MIS of the learning procedure is asymptotically 1.

Quantized models. When the parameters space is quantized, meaning that Θ is discrete, usually finite, it is possible to state high security results as well. Assume that $\mathcal{A}$ minimizes $L_n : \theta \mapsto \frac{1}{n} \sum_{i=1}^{n} l_\theta(x_i, y_i)$ over a quantized parameters space Θ, where $l_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ is a loss function and $z \coloneqq (x, y)$. Besides, it is common practice to adapt the size of the architecture with respect to the number of samples in the training dataset. Consequently, we suppose that the architecture follows a scaling law with respect to the number of samples $\Theta \coloneqq \Theta_n$. Then the following result holds,

Theorem 5 (Aubinais et al. (2025)) *Let* $\theta^* \in \Theta$ *be such that* $\mathbb{E}[l_{\theta^*}(x, y)] = \min_{\theta \in \Theta} \mathbb{E}[l_\theta(x, y)]$ *and we let* $m_n = \min\{\mathbb{E}[l_\theta(x, y)] - \mathbb{E}[l_{\theta^*}(x, y)] : \theta \neq \theta_*\}$. Then if $m_n\sqrt{n} \underset{n\to\infty}{\to} \infty$ and $m_2^n \underset{n\to\infty}{\to} 0$, then there exists a constant $c_{\ell,P}$ such that

$$\lim_{n\to\infty} -\frac{1}{n(m_2^n)^2} \log\left(1 - \mathrm{Sec}_{\nu,\lambda,n}(P, \mathcal{A})\right) \geq c_{\ell,P}.$$

3 Discussion

The hypothesis that the learning procedure can be expressed as a function of the empirical distribution is satisfied for any learning procedure minimizing an empirical cost. In particular, maximum likelihood based algorithms or Bayesian methods are special cases.

The choice of measure made in this article imply that the MIS is a measure of security on the learning procedure. One could understandably want to consider the security of a trained model, by changing the definition of accuracy and considering the quantity $\sup_\phi d\left(\phi; z_1, \cdots, z_n\right)$ conditionally to the training dataset $\{z_1, \cdots, z_n\}$, for some metric d, where d scales from 0 to 1. Natural choices of d include $d(\phi; z_1, \cdots, z_n) = \mathbb{P}(\phi(\hat{\theta}_n, \tilde{z}) = 1 - T \mid z_1, \cdots, z_n)$ or $d(\phi; z_1, \cdots, z_n) = TPR(\phi; z_1, \cdots, z_n)$. Unfortunately, this problem is degenerate as for those two choices of metric d, it can be easily shown that,

$$\sup_\phi d\left(\phi; z_1, \cdots, z_n\right) = 1,$$

for any learning procedure and any dataset. In other words, the underlying problem of studying the best performing MIA conditionally to the data has no relevant insight.
Theorem 4 and 5 respectively suggest that discretizing either the data or the parameter space enhance the security.

Bibliography

Aubinais, E., Gassiat, E., Piantanida, P. (2024), Fundamental Limits of Membership Inference Attacks on Machine Learning Models, *arXiv preprint arXiv:2310.13786*

Aubinais, E., Formont, P., Piantanida, P., Gassiat, E. (2025), Membership Inference Risks in Quantized Models: A Theoretical and Empirical Study, *arXiv preprint arXiv:2502.06567*

Del Grosso, G., Pichler, G., Palamidessi, C. and Piantanida, P. (2023), Bounding information leakage in machine learning, *Neurocomputing, 534,1 17.*

Dwork, C., Roth, A., et al. (2014), *The algorithmic foundations of differential privacy, volume 9.*

Hartley, J. and Tsaftaris, S. A. (2022), Measuring unintended memorisation of unique private features in neural networks, *arXiv preprint arXiv:2202.08099.*

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S. and Zhang, X. (2022), Membership inference attacks on machine learning models: A survey, *ACM Computing Surveyx (CSUR), 54(11s):1-37*

Shokri, R., Stronati, M. Song, C. and Shmatikov, V. (2017), Membership inference attacks against machine learning models, *IEEE symposium on security and privacy (SP), pages 3-18.*

IA, DURABILITÉ ET DÉVELOPPEMENT RÉGIONAL : UNE APPROCHE PAR CLUSTERS DES TRANSITIONS NUMÉRIQUE ET ÉCOLOGIQUE

Gianmarco Borrata & Simona Cafieri[2]

[1]*Università Federico II, Italie, gianmarco.borrata@unicampania.it*
[2]*ISTAT, Institut national de Statistique, Italie, simona.cafieri@istat.it*

Résumé. Ces dernières années, l'interaction entre la numérisation et la préservation de l'environnement est apparue comme un moteur essentiel du développement durable et inclusif dans les régions européennes. Ce travail explore l'intersection de l'intelligence artificielle (IA), en particulier l'IA générative, et les stratégies de durabilité environnementale, en analysant leur potentiel combiné pour remodeler les économies régionales et améliorer la résilience.
En employant une méthode Clusterwise au niveau européen NUTS2, la recherche identifie des groupes de territoires en fonction de leur degré de préparation au numérique et de leur engagement en faveur de la transition écologique. L'étude met particulièrement l'accent sur le rôle de l'IA générative dans l'accélération de l'innovation, l'optimisation de l'efficacité des ressources et la promotion de l'élaboration de politiques adaptatives.
Cette approche permet de mieux comprendre comment les contextes régionaux, les cadres politiques et les investissements dans les infrastructures interagissent pour produire divers résultats. Alors que certaines régions tirent parti de l'IA pour favoriser une transformation durable - grâce à la modélisation environnementale prédictive, à la gestion intelligente des ressources et aux économies circulaires renforcées par l'IA - d'autres se heurtent à des obstacles dus à des lacunes dans l'infrastructure numérique ou à des contraintes réglementaires, ce qui exacerbe les disparités sociales et économiques existantes.
Les résultats mettent en évidence la complexité de l'équilibre entre l'efficacité économique, l'équité territoriale et les objectifs environnementaux, démontrant qu'une stratégie uniforme a peu de chances d'aboutir à une durabilité à grande échelle. Au contraire, la recherche souligne la nécessité d'interventions adaptées au contexte qui alignent les capacités de l'IA sur les besoins régionaux.
En intégrant des outils analytiques avancés à des stratégies politiques tournées vers l'avenir, cette étude fournit aux décideurs politiques et aux parties prenantes une feuille de route fondée sur des données. Elle montre comment les régions européennes peuvent exploiter l'IA pour favoriser la croissance économique, l'inclusion sociale et la résilience écologique, en veillant à ce que les transitions numériques et environnementales se renforcent mutuellement au lieu de creuser les inégalités existantes.

Mots-clés. Développement régional, durabilité, IA, Clusterwise

Abstract. In recent years, the interplay between digitalization and ecological preservation has emerged as a crucial driver of sustainable and inclusive development across European regions. This work explores the intersection of artificial intelligence (AI), particularly generative AI, and environmental sustainability strategies, analyzing their combined potential to reshape

regional economies and enhance resilience.

By employing a Clusterwise method at the NUTS2 European level, the research identifies groups of territories based on their degree of digital readiness and commitment to ecological transition. The study places special emphasis on the role of generative AI in accelerating innovation, optimizing resource efficiency, and fostering adaptive policymaking.

This approach enables a deeper understanding of how regional contexts, policy frameworks, and infrastructural investments interact to produce diverse outcomes. While some regions leverage AI to drive sustainable transformation—through predictive environmental modeling, smart resource management, and AI-enhanced circular economies—others face barriers due to digital infrastructure gaps or regulatory constraints, exacerbating existing social and economic disparities.

The findings highlight the complexity of balancing economic efficiency, territorial equity, and environmental goals, demonstrating that a uniform strategy is unlikely to yield broad-scale sustainability. Instead, the research underscores the need for context-sensitive interventions that align AI capabilities with regional needs.

By integrating advanced analytical tools with forward-looking policy strategies, this study provides a data-driven roadmap for policymakers and stakeholders. It showcases how European regions can harness AI to foster economic growth, social inclusion, and ecological resilience, ensuring that digital and environmental transitions reinforce one another rather than widen existing inequalities.

Keywords. Regional development, sustainability, AI, Clusterwise

1. Introduction

Ces dernières années, l'intelligence artificielle générative (GenAI) a pris un rôle central dans les processus d'innovation technologique. Parallèlement, la transition écologique est devenue une priorité croissante dans les agendas politiques mondiaux. L'intégration de la GenAI dans les stratégies de durabilité environnementale offre des opportunités sans précédent pour l'optimisation des ressources, la réduction de l'impact environnemental et le soutien aux décisions politiques basées sur les données. Cependant, l'efficacité de cette intégration varie significativement entre les régions européennes en raison des différences dans les infrastructures numériques et les niveaux d'adoption technologique. Dans cette étude, nous appliquons des techniques avancées de régression et de clustering pour analyser l'interaction entre la numérisation, l'adoption de l'IA générative et la durabilité environnementale dans les régions européennes au niveau NUTS2.

2. Méthodologie

Pour identifier et classifier les régions européennes selon leur maturité numérique et leur engagement écologique, nous utilisons une combinaison de techniques de clustering et de régression avancée. Notre approche s'articule selon les étapes suivantes :

2.1 Collecte et Prétraitement des Données

Nous avons collecté des données provenant d'Eurostat, de l'OCDE et de la Banque Mondiale, en nous concentrant sur des indicateurs clés :

- **Maturité numérique :** accès au haut débit, investissements en R&D sur l'IA, compétences numériques.
- **Engagement écologique :** réduction des émissions de CO_2, part des énergies renouvelables, économie circulaire.
- **Adoption de la GenAI :** nombre de startups IA par région, politiques d'innovation.

Les données ont été normalisées à l'aide de la transformation Min-Max pour assurer leur comparabilité entre les régions.

2.2 Clustering des Régions Européennes

Nous avons appliqué la **Clusterwise Regression (CR)** pour segmenter les régions en groupes homogènes, en combinant :

- **Clustering K-Means** avec la distance euclidienne pour une première segmentation.
- **Algorithme de clustering hiérarchique** pour affiner les groupes et vérifier la robustesse des résultats.

Les clusters résultants identifient des régions avec **un niveau élevé, moyen ou faible de numérisation et de durabilité**.

2.3 Modèles de Régression pour la Prévision de l'Impact de la GenAI sur la Durabilité

Nous utilisons la **régression fonctionnelle Clusterwise LDQ (Logarithm-Derivative Quantile)** pour modéliser la relation entre l'adoption de la GenAI et les indicateurs de durabilité, en identifiant les variables clés influençant la transformation écologique.

3. Résultats et Discussion

3.1 Analyse des Clusters

Les résultats montrent trois macro-clusters :

- **Cluster 1 : Régions à forte numérisation et durabilité avancée** (Europe du Nord, Pays-Bas, certaines régions allemandes).
- **Cluster 2 : Régions avec une numérisation intermédiaire et un engagement écologique moyen** (Italie du Nord, Espagne, France).
- **Cluster 3 : Régions avec une numérisation limitée et une durabilité réduite** (Europe de l'Est, Sud de l'Italie, Grèce).

3.2 Régression et Prévisions

L'analyse régressive met en évidence que **l'adoption de la GenAI a un impact positif sur la réduction des émissions et l'efficacité des ressources**, avec des effets plus marqués dans les régions déjà avancées numériquement. Toutefois, dans les régions moins développées sur le plan numérique, l'impact est limité en raison d'infrastructures insuffisantes et de faibles investissements en R&D.

4. Conclusions et Implications Politiques

Les résultats indiquent qu'**une approche uniforme de la numérisation et de la durabilité n'est pas efficace**. Une stratégie différenciée est nécessaire :

- **Cluster 1 :** Encourager le développement d'applications GenAI pour la durabilité.
- **Cluster 2 :** Renforcer la numérisation pour accélérer l'impact de la GenAI.
- **Cluster 3 :** Investir dans les infrastructures numériques et la formation pour combler l'écart technologique.

Cette analyse fournit une feuille de route basée sur les données pour l'adoption de l'IA générative dans la transition écologique européenne.

Bibliographie

Da Silva R. A. and de Carvalho F. d. A., (2021), Weighted clusterwise linear regression based on adaptive quadratic form distance, *Expert Systems with Applications,* vol. 185, p. 115609.

Openshaw S. (1977), A geographical solution to scale and aggregation problems in region building, partitioning and spatial modelling, *Transactions of the institute of british geographers*, pp. 459–472,.

Preda C. and Saporta G.(2005), Clusterwise pls regression on a stochastic process, *Com putational Statistics & Data Analysis*, vol. 49, no. 1, pp. 99–108.

Zhenzhen L,(2024) The impact of artificial intelligence technology innovation on economic development -- from the perspective of generative AI products, , *Journal of Education, Humanities and Social Sciences*, vol 27

Fairness is in the details: Face Dataset Auditing

Valentin Lafargue [1,2,3] & Emmanuelle Claeys [4] & Jean-Michel Loubes [2,3]

[1] *IMT, France et {valentin.lafargue}@univ-tlse3.fr* [2] *ANITI, France* [3] *INRIA, France et {jean-michel.a.loubes}@inria.fr* [4] *IRIT, France et Emmanuelle.Claeys@irit.fr*

Résumé. Auditer consiste à vérifier une conformité selon une politique donnée. Dans le cadre de l'AI Act de l'Union Européenne, il s'agit de l'application des règles de *fairness*, d'équité et de transparence. Cela nécessite d'éviter les biais présents dans l'entraînement d'un modèle d'apprentissage qui entraînerait une discrimination pour un sous groupe d'individus lorsque ce modèle est mis en production. L'évaluation du biais sur des *datasets* d'images est une tâche complexe, car elle nécessite préalablement une étape d'extraction des *features*, puis de considérer la qualité de cette extraction dans la statistique de test. Cet article propose une méthodologie robuste pour l'audit de *datasets* d'images sur des attributs dits "sensibles" tels que le genre, l'âge et l'ethnicité. Cette méthodologie comprend à la fois une étape d'extraction des attributs et une analyse statistique. La première étape inclut une nouvelle architecture de réseau de neurones convolutifs (CNN) dédiée à l'extraction des attributs sensibles avec un nombre restreint d'annotations manuelles. La seconde étape compare les distributions des attributs sensibles entre les sous-groupes par un nouveau test statistique qui inclut l'imprécision du modèle d'extraction. Notre pipeline constitue une nouvelle méthodologie complète et automatisée pour l'audit des jeux de données d'images. Nous illustrons notre méthode avec deux *datasets* annotés manuellement.

Mots-clés. Audit, *Datasets*, Bias, Images, Test statistique, Classification, Fitzpatrick

Abstract. Auditing involves verifying the proper implementation of a given policy. As such, auditing is essential for ensuring compliance with the principles of fairness, equity, and transparency mandated by the European Union's AI Act. Moreover, biases present during the training phase of a learning system can persist in the modeling process and result in discrimination against certain subgroups of individuals when the model is deployed in production. Assessing bias in image datasets is a particularly complex task, as it first requires a feature extraction step, then to consider the extraction's quality in the statistical tests. This paper proposes a robust methodology for auditing image datasets based on so-called "sensitive" attributes, such as gender, age, and ethnicity. The proposed methodology consists of both an attribute extraction phase and a statistical analysis phase. The first phase introduces a novel convolutional neural network (CNN) architecture specifically designed for extracting sensitive attributes with a limited number of manual annotations. The second phase compares the distributions of sensitive attributes across subgroups using a novel statistical test that accounts for the imprecision of the attribute extraction model. Our pipeline constitutes a comprehensive and fully automated methodology for dataset auditing. We illustrate our approach using two manually annotated datasets.

Keywords. Audit, Dataset, Bias, Images, Statistical test, Classification, Fitzpatrick

1 Introduction

L'industrialisation massive des systèmes d'apprentissage automatique a amplifié les préoccupations concernant l'équité, la transparence et la responsabilité. L'audit des *datasets* utilisés pour l'apprentissage est devenu une étape préalable à l'entraînement du modèle, et devient un enjeu juridique dans le cadre de la réglementation imposée par l'AI Act. Pour un *dataset* constitué d'images, il est nécessaire de déterminer quelles variables considérer pour évaluer les biais. Cette tâche est complexe car chaque image n'a pas de formalisation des attributs "sensibles" explicites : le sexe, l'âge et le genre par exemple, ne sont que visuellement déduits de l'image, mais pas directement encodés par un attribut. L'extraction de ces attributs par une annotation manuelle devient inenvisageable si le nombre de données est important. Pour extraire automatiquement ces attributs, un modèle prédictif de type réseau convolutif peut être utilisé, bien qu'il requière également une annotation manuelle (moindre), nécessaire à son entraînement. Une fois entraîné, le modèle peut prédire les attributs sensibles et annoter complètement le *dataset* (avec un taux d'erreur relatif au modèle). Nous définissons dans notre contexte le biais dans un *dataset* d'images comme des différences statistiquement significatives de représentations (par exemple, une sous-représentation d'une catégorie ethnique et/ou genrée). Les statistiques de tests ne considèrent généralement pas l'incertitude autour de l'annotation (ou peuvent être difficiles à utiliser en multidimensionnel). Nous proposons d'évaluer les biais en comparant directement les répartitions des attributs sensibles de sous-groupes, tout en considérant l'imprécision du modèle d'extraction des attributs. La section 2 présente un état de l'art sur les méthodes d'extraction d'attributs sensibles et les tests statistiques robustes aux erreurs de labellisation. La section 3 présente notre méthodologie, La section 4 explicite les résultats obtenus sur deux *datasets*, puis la section 5 donne des perspectives sur des travaux futurs.

2 Littérature

L'extraction d'attributs dits 'sensibles' à partir d'une image, nécessite premièrement de choisir un type d'attributs, puis une méthode d'extraction automatisée.

Choix des attributs Contrairement à la classification du sexe des individus (nous considérons ici une classification binaire du sexe), la classification ethnique est un choix arbitraire [9]. De nombreux datasets académiques de référence utilisent la classification du US Census 2000 qui sépare les individus en cinq classes (*Latino, Asiatique, Blanc, Noir, Autre*). Une alternative à cette classification est d'utiliser une classification issue de la dermatologie. Par exemple, l'ITA (Individual Typology Angle) [18, 14] estime des nuances continues de peau selon la luminosité. La classification Fitzpatrick [6] classe les individus en couplant leur réaction au soleil avec des informations complémentaires (cheveux, couleur des yeux, ...). Une fois le type d'attribut statué, leur extraction à partir des images doit être réalisée.

Extraction Selon la taille du *dataset*, l'extraction manuelle n'est pas toujours envisageable, et requiert donc l'utilisation d'un classifieur pour annoter automatiquement une partie du *dataset*. Les CNN sont particulièrement adaptés aux images pour extraire des attributs identifiables visuellement. Pour identifier des attributs liés à l'ethnicité, au genre ou l'âge, le CNN doit capturer des nuances de peau [5] mais également des informations complémentaires (par exemple la texture des cheveux) [17]. Le choix de l'architecture doit donc permettre d'identifier les régions de l'image où ces informations sont présentes, et exclure les régions sans intérêt comme l'arrière-plan. Parmi les méthodes existantes, l'architecture FairFace [12] propose de détecter le visage puis de classifier l'âge et le genre à l'aide d'une architecture ResNet34 [10]. Des méthodes de suppression d'arrière-plan sont proposées dans [7]. La méthode de [16] utilise une architecture de U-net imbriqués appelée U^2-net. Enfin, l'isolement de la région de la peau segmentée peut être réalisé par DeepLabv3 [4] en utilisant un modèle MobileNetV3 Large Backbone [11]. Cependant, ces méthodes ne proposent pas d'étude sur la taille du *dataset* d'annotation, ni de paramétrage spécifique à l'extraction de l'ITA ou de la classification Fitzpatrick.

Comparaison des distributions Une fois les attributs extraits des images, l'étude de leur distribution doit être réalisée afin de vérifier la présence de biais. Considérons un *dataset* $\mathcal{D}$, avec d attributs : $X^0, \ldots, X^d$. Soit un attribut X^j choisi pour obtenir K partitions selon ses K modalités. Nous souhaitons observer la distribution de X^0 en fonction de X^j : $D(X^0|X^j)$. La distance de Wasserstein permet de quantifier la distance entre deux distributions sur des échantillons multidimensionnels. Elle est donnée par:

$$W_p(\mu,\nu) = \big(\inf_{\pi\in\Gamma(\mu,\nu)} \int_{M\times M} d(x,y)^p d\pi(x,y)\big)^{1/p}$$

où ici $p \geq 1$, W_p est la p^{th} distance de Wasserstein, $\Gamma(\mu,\nu)$ est la distribution jointe totale π avec la distribution marginale μ et ν. La distance entre deux points x et y dans M est donnée par $d()$. À partir de cette distance, un test à deux échantillons peut être réalisé. Il existe de nombreux tests alternatifs permettant de comparer les distributions de deux modalités d'attribut en fonction de leur type. Dans notre contexte, l'attribut peut-être de type continu (comme l'ITA), binaire (comme le genre) ou multimodal (comme la classification Fitzpatrick). Le choix du test dépendra des hypothèses sur la distribution de l'attribut. En fonction du choix et des hypothèses, le test du χ^2, le test des moyennes basé sur le TCL, le test de Kolmogorov-Smirnov peuvent par exemple être utilisés. Cependant, aucune des méthodes précédemment citées ne prend en compte l'erreur du classifieur qui permet l'extraction des attributs. La section suivante présente notre contribution en ce sens, puis nous exposerons les résultats sur deux *datasets* académiques.

3 Contribution

Nous procédons à une labellisation manuelle (classification Fitzpatrick) de deux *datasets* académiques pour pouvoir évaluer la taille minimale nécessaire à l'entraînement. L'ITA est

Accuracy (%)		
Gender	Men (M) : 97 % →3% permutation with W	Women (W) : 95 % → 5% permutation with M
Fitzpatrick class V given the gender	V knowing predicted M : 93 % → 7% remplacement $x \sim \mathcal{U}(X_V \mid X_G = M,\ X_{pred} = False)$	V knowing predicted W : 85 % →15% remplacement $x \sim \mathcal{U}(X_V \mid X_G = W,\ X_{pred} = False)$

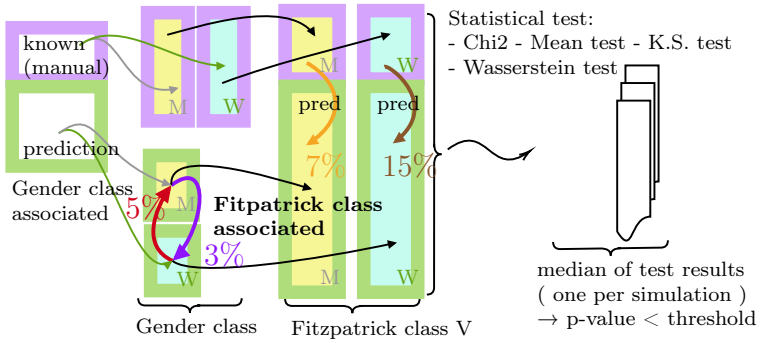


Figure 1: Diagramme expliquant le processus d'agrégation des tests statistiques d'égalité de répartition prenant en compte l'imprécision du modèle. Ce schéma montre la minimisation de l'impact des erreurs sur le test en utilisant les permutations et les remplacements adaptés.

calculé par :

$$\text{ITA} = \arctan\left(\frac{L* - 50}{b*}\right) \times \frac{180}{\pi} \tag{1}$$

Où $L*$ et $b*$ correspondent à des variables décrivant la lumière et la couleur du pixel moyen associé à la peau du visage. Pour classifier nos attributs, nous entraînons un CNN de type ResNet50 ou Resnet101 couplé avec un *overhead network*. Trois *inputs* sont comparés : (1) l'utilisation des images avec arrière-plan (soit l'image originale) (2) l'image sans arrière-plan, et (3) la peau de l'individu (isolée par segmentation). L'ITA (Individual Typology Angle) est ajouté comme information complémentaire à l'image. Nous avons évalué deux couches d'*embedding* pour l'extraction : (1) la sortie de la couche finale (1 000 dimensions), et (2) l'avant-dernière couche (2 048 dimensions) (Figure 4 en annexe). Nous faisons varier la taille de l'ensemble d'entraînement pour chercher à limiter l'annotation manuelle tout en considérant la précision du modèle. Pour étudier la présence de biais dans les images annotées automatiquement, nous proposons d'intégrer la précision du modèle dans les tests du χ^2, des moyennes et le test de la distance de Wasserstein ($p = 2$). Ces tests sont modifiés suivant la méthodologie suivante :

1. Nous calculons la précision du modèle (jeu de validation) P^{X^0} sur l'attribut X^0.
2. Nous permutons aléatoirement $100 \times (1 - P^{X^0})$ % des labels associés à X^0 sur les données non annotées manuellement. Les autres labels restent inchangés.

Dans le cas d'un test de parité, nous comparons la proportion estimée avec un paramètre de référence (c'est-à-dire avec des valeurs simulées selon ce paramètre). À noter que les variables multimodales catégorielles sont traitées comme des variables binaires en un-contre-tous. Pour un test de répartition, une variable conditionnelle X^j, $j \neq 0$ ayant K modalités, est utilisée

pour partitionner le *dataset* en K partitions. Soit x_i^j une modalité particulière de X^j telle que $X^j \in \{x_1^j, \dots x_K^j\}$. Soit X^0 la variable étudiée, ayant K' modalités, avec $x_{i'}^0$ une modalité particulière de X^0 telle que $X^0 \in \{x_1^0, \dots x_{K'}^0\}$. Nous calculons $P_{i,i'}^{X^0,X^j} = \{\mathbb{P}[\hat{x}_{i'}^0 = x_{i'}^0 | x_i^j], i \in [\![1, K]\!]\}$ la précision de notre modèle de classification pour la modalité $x_{i'}^0$ conditionnée à x_i^j. Pour chaque modalité x_i^j nous remplaçons aléatoirement $100 \times (1 - P_{i,i'}^{X^0,X^j})\%$ des variables $X^0 \in \{X | X^j = x_i^j, X^0 = x_{i'}^0\}$ annotées automatiquement, par des variables $X^0 \in \{X | X^j = x_i^j, X^0 = x_{i'}^0\}$ annotées manuellement.

4 Résultats

Notre objectif est, en tenant compte des erreurs de classification, (1) d'auditer s'il existe une différence significative entre les proportions observées à travers trois attributs sensibles : le genre, l'âge et les classifications Fitzpatrick; (2) d'auditer s'il existe une différence significative entre les proportions observées sur les âges et la classification Fitzpatrick, **conditionnées au genre**. Nos résultats ont montré qu'une architecture par *embedding* ayant un espace latent de grande dimension, l'ajout de l'ITA au modèle et enfin le retrait de l'arrière-plan de l'image améliorent significativement l'apprentissage. Cependant, l'utilisation d'un masque de peau détériore les résultats. Cela peut être expliqué par la suppression de caractéristiques comme la couleur des cheveux, essentielles à la classification Fitzpatrick (Table 1 en annexe). L'ajout de l'ITA a montré une amélioration notable pour des petits jeux d'entraînements avec une taille d'embedding de dimension 1000. Nous pensons que les caractéristiques liées à la peau sont suffisamment capturées pour l'*embedding* de taille 2048 mais perdues lors de la réduction de dimension (*embedding* de taille 1000).

Nous comparons les hypothèses acceptées en fonction du *sample size* pour les deux *datasets*. Pour le *dataset GAN*, à partir d'un *sample size*$=$ 1000 le test de parité (1) n'a abouti qu'à un seul faux rejet sur les 15 hypothèses testées (l'erreur concerne la catégorie Fitzpatrick V selon un paramètre de $\frac{1}{5}$). Toujours pour un paramètre de *sample size*$=$ 1000, notre test d'égalité de répartition (2) a produit trois faux rejets sur les 14 tests. Aucun faux positif n'a été recensé sur ce *dataset*. Pour le *dataset CelebA*, nous avons obtenu deux faux négatifs et un faux positif pour le test de parité (1). Pour le test de répartition (2) l'intégralité des tests ont abouti à des conclusions conformes à l'annotation manuelle complète. Pour un *sample size*$\geq$ 1000 (qui apparaît comme un paramétrage réaliste), notre audit révèle que, pour le *dataset GAN*, les proportions observées sur l'attribut *age* ne sont par représentatives des proportions recensées dans la population. La répartition des âges selon le genre est considérée comme uniforme. Cependant, il existe un fort biais lié au genre, par exemple si l'on est une femme, la probabilité d'être dans la classe Fitzpatrick I est de 10% contre 7% pour un homme. A contrario, pour une femme, la probabilité d'être dans la classe Fitzpatrick VI est de 9% contre 12% pour un homme (Table 5 en annexe). Pour le *dataset CelebA*, un constat similaire pour la parité peut être fait, bien que la population d'âgée de 40 à 59 ans soit représentative du paramètre recensé. Il existe également un biais important lié au sexe dans ce dataset, par exemple, la tranche d'âge 20-29 est sur-représenté chez les femmes.

5 Conclusion & Discussion

Nous avons proposé une nouvelle méthode d'audit du biais, limitant l'annotation manuelle des images (avec un seuil à 1000 annotations) et robuste aux erreurs d'annotation automatisées. Nous soulignons l'importance d'évaluer la représentativité des individus avant d'utiliser une base de données d'images pour de l'apprentissage automatique, afin de limiter les éventuelles discriminations. Dans le cadre de l'AI Act, qui exigera des entreprises qu'elles attestent la conformité des données d'entraînement pour des modèles de machine learning (ML), nous espérons que cet audit sera un premier outil à leur disposition. Nos travaux futurs viseront à contrôler la présence de biais en mode *online* dans les modèles génératifs.

References

[1] *Our World in Data*. 2022. URL: https://ourworldindata.org/.

[2] *World Bank Data*. 2022. URL: https://data.worldbank.org/.

[3] Vishnu Naresh Boddeti, Gautam Sreekumar, and Arun Ross. *On the Biometric Capacity of Generative Face Models*. 2023. arXiv: 2308.02065 [cs.CV].

[4] Liang-Chieh Chen et al. *Rethinking Atrous Convolution for Semantic Image Segmentation*. 2017. arXiv: 1706.05587 [cs.CV].

[5] Gianluigi Ciocca, Raimondo Schettini, and Silvia Zuffi. *A survey on methods for colour image indexing and retrieval in image databases*. 2002.

[6] Thomas B. Fitzpatrick. "The Validity and Practicality of Sun-Reactive Skin Types I Through VI". In: *Arch Dermatol.* (June 1988).

[7] Daniel Gatis. *Rembg*. 2022. URL: https://github.com/danielgatis/rembg.

[8] Generated, photos, and team. *Generated Dataset*. 2024. URL: https://generated.photos/solutions/academic-research.

[9] Alex Hanna et al. "Towards a critical race methodology in algorithmic fairness". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 501–512. ISBN: 9781450369367.

[10] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[11] Andrew Howard et al. *Searching for MobileNetV3*. 2019. arXiv: 1905.02244 [cs.CV]. URL: https://arxiv.org/abs/1905.02244.

[12] Davis E. King. *Max-Margin Object Detection*. 2015. arXiv: 1502.00046 [cs.CV].

[13] Yang Z. Liu J Bitsue HK. *Skin colour: A window into human phenotypic evolution and environmental adaptation*. 2024.

[14] Michele Merler et al. *Diversity in Faces*. 2019. arXiv: 1901.10436 [cs.CV].

[15] E. J. Parra, R. A. Kittles, and M. D. Shriver. "Implications of correlations between skin color and genetic ancestry for biomedical research". English (US). In: *Nature Genetics* 36.11 SUPPL. 1 (Nov. 2004). Copyright: Copyright 2008 Elsevier B.V., All rights reserved., S54–S60. ISSN: 1061-4036.

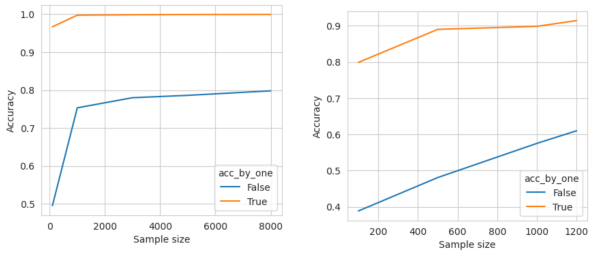
[16] Xuebin Qin et al. "U2-Net: Going deeper with nested U-structure for salient object detection". In: *Pattern Recognition* 106 (2020), p. 107404. ISSN: 0031-3203.

[17] William Thong, Przemyslaw Joniak, and Alice Xiang. *Beyond Skin Tone: A Multidimensional Measure of Apparent Skin Color*. 2023. arXiv: 2309.05148 [cs.CV].

[18] Yue Wu, Toshiyuki Tanaka, and Makio Akimoto. "Utilization of individual typology angle (ITA) and hue angle in the measurement of skin color on images". In: *bioimages* 28 (2020), pp. 1–8.

Annexe

Figure 2: Fitzpatrick classification (5 classes annotées manuellement), de gauche à droite, les phototypes I, II, III-IV, V, VI. La première ligne représente le jeu de donnée généré par un Style-GAN [8] [3] (10 000 images), la deuxième ligne est issue d'un sous-échantillon de 1500 images de CelebA. Nous mettons à disposition notre annotation manuelle à la communauté.



(a) GAN dataset (b) CelebA dataset

Figure 3: Impact de la taille de l'échantillon sur la précision du réseau de neurone.

Hyperparamètres	GAN Dataset	CelebA Dataset
Information sur la peau (ITA..)	+0.6	+0.2
Choix du modèle	+3.7	+4.2
Mask sur la peau	-1.7	+1.2
Enlever l'arrière-plan	+0.1	+1.8
Taille de l'espace latent (2048)	+1.9	+2.7

Table 1: Impact des hyperparamètres et de l'achitecture sur la précision du réseau de neurone classifiant l'échelle de Fitzpatrick.

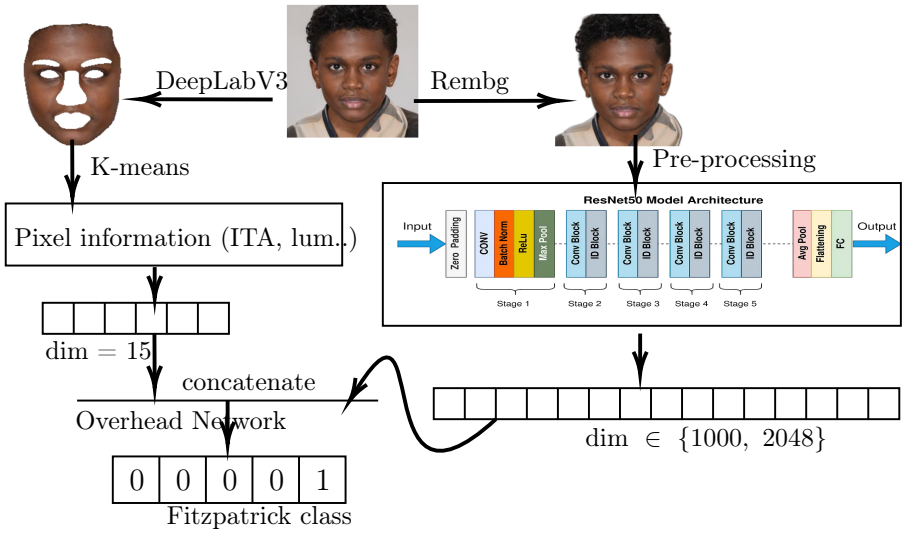


Figure 4: Pipeline for Fitzpatrick classification

Table 2: Test statistique sur la parité du genre ($X \sim \mathcal{B}(\frac{1}{2})$), de la classe de Fitzpatrick ($X \sim \mathcal{B}(\frac{1}{5})$), et de l'âge ($X \sim \text{RealDistr}_{\text{age}}$). ✓ signifie qu'aucun test n'a rejeté l'hypothèse que les distributions était issue d'une même loi, × signifie qu'au moins deux tests ont rejeté cette hypothèse. L'objectif est d'obtenir les mêmes conclusion que pour le paramétrage max *Sample size* .

(a) GAN

Sensitive Attributes	**Sample Size**				
	100	**500**	**1000**	**3000**	**8000**
Gender	×	×	×	×	×
Age					
0-2	✓	✓	✓	✓	✓
3-9	×	×	×	×	×
10-19	×	×	×	×	×
20-29	×	×	×	×	×
30-39	×	×	×	×	×
40-49	×	×	×	×	×
50-59	×	×	×	×	×
60-69	×	×	×	×	×
70+	×	×	×	×	×
Fitz. Class					
I	×	×	×	×	×
II	×	×	×	×	×
III-IV	×	×	×	×	×
V	×	×	×	×	✓
VI	×	×	×	×	×

(b) CelebA

Sample Size			
100	**500**	**1000**	**1200**
×	×	×	×
✓	✓	✓	✓
×	×	×	×
×	×	×	×
×	×	×	×
×	×	×	×
✓	✓	✓	✓
✓	✓	✓	✓
×	×	×	×
×	×	×	×
×	×	×	×
×	×	×	×
×	×	×	✓
×	×	×	✓
×	×	✓	×

Table 3: Test statistique sur l'égalité de répartition pour les classes de Fitzpatrick et l'âge, distributions étudiées après division suivant la variable de genre. La répartion de chaque attributs est comparée avec la distributions des attributs restants (*one versus all*). L'objectif est d'obtenir les mêmes conclusion que pour le paramétrage max *Sample size* .

(a) GAN

Sensitive Attributes	**Sample Size**				
	100	**500**	**1000**	**3000**	**8000**
Age					
0-2	✓	✓	✓	✓	✓
3-9	✓	✓	×	✓	✓
10-19	✓	✓	✓	✓	✓
20-29	$\checkmark_{2/3}$	✓	×	✓	$\checkmark_{2/3}$
30-39	✓	✓	✓	✓	✓
40-49	✓	✓	✓	✓	✓
50-59	✓	✓	✓	✓	✓
60-69	✓	✓	✓	✓	✓
70+	✓	✓	✓	✓	✓
Fitz. Class					
I	$\checkmark_{2/3}$	$\checkmark_{2/3}$	×	×	×
II	×	×	$\checkmark_{2/3}$	$\checkmark_{2/3}$	✓
III-IV	×	×	×	×	$\checkmark_{2/3}$
V	×	×	✓	×	$\checkmark_{2/3}$
VI	×	×	×	×	×

(b) CelebA

Sample Size			
100	**500**	**1000**	**1200**
✓	✓	✓	✓
✓	✓	✓	✓
✓	✓	✓	✓
×	×	×	×
✓	$\checkmark_{2/3}$	✓	✓
✓	✓	✓	✓
✓	✓	✓	✓
✓	✓	✓	✓
✓	✓	✓	✓
$\checkmark_{2/3}$	$\checkmark_{2/3}$	×	×
×	×	×	×
✓	✓	✓	✓
×	×	$\checkmark_{2/3}$	✓
✓	×	×	×

Table 4: Répartition des classes de l'âge en fonction du genre (en pourcentage %). Nous avons mis en gras ce qui a été détecté comme une différence significative par notre pipeline d'audit.La répartition de l'âge réelle fut recensée par [2][1]

Tranche d'âge	**Gan (10 000)**				**CelebA (1 500)**				**Réelle**
	Femme	**Homme**	**Diff**	**Total**	**Femme**	**Homme**	**Diff**	**Total**	**Total**
0-2	0	0	0	0.3	0	0	0	0	2
3-9	3	6	2	4	1	0	0	0.3	12
10-19	5	8	4	6	2	4	2	3	17
20-29	69	50	19	61	**73**	**36**	**37**	48	17
30-39	19	25	6	22	19	29	9	26	15
40-49	2	6	4	4	3	14	11	11	13
50-59	1	3	2	2	2	10	8	7	11
60-69	0	1	0	0.3	0	6	6	4	8
70+	0	0	0	0.05	0	6	6	0.1	7

Table 5: Répartition des classes de Fitzpatrick en fonction du genre (en pourcentage %) à partir d'annotation manuelle. Nous avons mis en gras ce qui a été considéré comme une différence significative par notre pipeline d'audit (*sample size* maximale). La répartition réelle fut recensée par [15] [13]. H : Homme, F : Femme, T : Total, Dif : $|H - F|$, Max/Min : $\frac{max(H,F)}{min(H,F)}$

Classe de Fitzpatrick	**Gan (10 000)**					**CelebA (1 500)**					Réelle
	F	**H**	**Diff**	**Max/Min**	**T**	**F**	**H**	**Diff**	**Max/Min**	**T**	**T**
I	**10**	**7**	3	**1.37**	**2**	2	**31**	**29**	**12.5**	22	4
II	32	36	4	1.12	34	**66**	**52**	**14**	1.28	56	15
III - IV	35	30	5	1.18	33	15	7	7	2.03	10	50
V	15	16	1	1.04	15	16	7	9	2.34	10	19
VI	**9**	**12**	3	**1.41**	10	**1**	**3**	2	**3.03**	2	12

ANONYMIZATION OF GEOREFERENCED HEALTH MICRODATA A GLIMPSE OF WORK IN PROGRESS

Rainer Lenz[1] & Lydia Jehmlich[1] & Simon Cremer[1]

[1] *Cologne university of technology, arts and sciences, 50679 Cologne, Germany*
{rainer.lenz, lydia.jehmlich, simon.cremer}@th-koeln.de

Abstract. Georeferenced microdata has extraordinary potential for research and teaching, public administration and business. Key questions about the future and sustainability of our society can only be answered with high-quality and accessible geodata. Based on the EU INSPIRE directive, the statistical offices of the EU are successively implementing the georeferencing of their surveys. Whether for the use of targeted advertising measures or tracing the spatial spread of viruses such as the recent corona virus: georeferenced microdata can - depending on the attributes it is provided with - hold enormous added value for society, science and research. However, the desired information can often not be extracted despite the inherent analytical content. The reason for this is that access to personal georeferenced datasets is severely restricted, as these are subject to statutory data protection. One way out of this dilemma is to apply a suitable anonymization method that guarantees data protection without significantly reducing the analytical validity of data. This paper discusses selected anonymization methods being most promising for anonymizing georeferenced health data for research purposes, as they offer scope for combinations or more specific adaptations in order to balance out the trade-off between security and analytical validity of georeferenced health microdata.

Keywords. Geoprivacy, location privacy, disclosure risk, health data, data anonymization, spatial analysis

Résumé. Les microdonnées géoréférencées présentent un potentiel extraordinaire pour la recherche et l'enseignement, l'administration publique et les entreprises. Seules des géodonnées de haute qualité et accessibles permettent de répondre à des questions essentielles sur l'avenir et la durabilité de notre société. Sur la base de la directive européenne INSPIRE, les instituts de statistique de l'UE mettent successivement en œuvre le géoréférencement de leurs enquêtes. Qu'il s'agisse d'utiliser des mesures publicitaires ciblées ou de retracer la propagation spatiale de virus tels que le récent virus corona, les microdonnées géoréférencées peuvent, selon les attributs dont elles sont dotées, apporter une énorme valeur ajoutée à la société, à la science et à la recherche. Cependant, les informations souhaitées ne peuvent souvent pas être extraites malgré le contenu analytique inhérent. La raison en est que l'accès aux ensembles de données géoréférencées personnelles est fortement limité, car elles sont soumises à la protection légale des données. Pour sortir de ce dilemme, il est possible d'appliquer une méthode d'anonymisation appropriée qui garantisse la protection des données sans réduire de manière significative la validité analytique des données. Ce document examine les méthodes d'anonymisation sélectionnées qui sont les plus prometteuses pour l'anonymisation des données de santé géoréférencées à des fins de recherche, car elles offrent la possibilité de combinaisons ou d'adaptations plus spécifiques afin d'équilibrer le compromis entre la sécurité et la validité analytique des microdonnées de santé géoréférencées.

1 Introduction

Anonymization of health data is a major challenge due to the high degree of individuality of personal and household data. According to the European General Data Protection Regulation (GDPR) Art. 9, these data are considered particularly sensitive and worthy of protection. Hence, the bulk of publications specifically on the anonymization of geodata deals with health data. To identify confidential information, a potential data attacker can use the geocoordinates of the registration addresses of patients, for example. Not only tables of numbers, but also characteristics such as image data (X-rays, CT scans), names of treating physicians or administered medication can contribute to the identification of individuals. What are called "health apps" nowadays collect data on the health status of their users. The rapid development of data availability and methodology, combined with growing computing power, is increasingly calling traditional methods of anonymization and statistical confidentiality into question. When applying anonymization methods, a bicriteria optimization problem must be solved. On the one hand, the analysis potential of the data should be maximized, while on the other the data should be protected as well as possible.

2 Statistical disclosure control

A fine regionalization of individual data harbors a great risk of unambiguous assignments. The high sensitivity of the data usually requires the restriction to high aggregation levels, in applied research with microdata sets outside of research data centers usually at the level of appropriate administrative districts. To address this problem, the development of realistic privacy metrics that can better quantify the likelihood of and loss from potential re-identification with respect to geospatial data is essential. According to the GDPR natural persons are protected by regulating the processing of data relating to an identified or identifiable natural person. Anonymization of data in such a way that the remaining available information does not allow any conclusions to be drawn about individual subjects, but still has sufficient information potential, is a core concern of every data producer. This is a basic principle of European statistics and aims to guarantee the confidentiality of data while at the same time maintaining its usability for statistical purposes (Eurostat 2023).

3 Anonymity and analytical validity of georeferenced data

The what is called *k*-anonymity provides information on the degree to which data sets can be re-identified by combining their quasi-identifiers. A dataset is *k*-anonymous if each sequence of values appears with at least *k* occurrences in the data. The concept of *k*-anonymity can also be applied to spatial datasets. This is then referred to as spatial *k*-anonymity, which is the most widely used metric for measuring anonymity in the masking of sensitive geodata (Broen et al. 2021). It basically describes the number of masked points that are closer or equally close to the original location as the masked original point itself. Thus, with a spatial *k*-anonymity, at least *k-1* masked points are closer to an original address than the associated masked point. As a result, the probability of a data attacker locating the original location is at most *1/k*. However, there is also criticism of k-anonymity. To counteract this, the concept of spatial k-anonymity can be extended with approaches such as l-diversity, t-closeness, or the concept of differential privacy (Dwork 2006).

Anonymization is always accompanied by a reduction of information. Either information is suppressed to such an extent that it can no longer be assigned or only with disproportionate

effort, or this protection is achieved by modifying the data to such an extent that the usefulness of it is (partially) lost. Either way, this means a loss of information. If a specific research objective is already being pursued with the data to be anonymized, the anonymization method can be specifically adapted to this objective. Typical questions regarding health data deal with the population's access to medical care (specialists, hospitals, pharmacies, etc.) or the classification of diagnoses (ICD-code). Accordingly, there are many studies on the application of location masking methods to health datasets such as leukemia (Armstrong et a.. 1999), lung cancer (Kwan et al. 2004) or, more recently, the Covid-19 (Houfaf-Khoufaf and Touya 2021). The lack of reporting uncertainty levels of spatial data makes analyses based on such datasets susceptible to bias.

4 Methods of data anonymization

This chapter presents pre-selected anonymization methods that serve as basic mechanisms for anonymizing georeferenced microdata. These methods have proven to be particularly suitable for further studies on anonymization of health data. First of all, the considered dataset is pseudonymized, which is in most cases not the last step to protect confidential data. In pseudonymization, direct identifiers such as name or address are replaced by a pseudonym, such as a sequential number or a code, to make it more difficult to correctly assign individuals. The further anonymization can then be roughly divided into two categories: aggregation and perturbation of quasi-identifiers and sensitive attributes. The following sections present methods from both areas, which are most promising for anonymizing health microdata. Nevertheless, no claim is made to completeness. It should also be mentioned that these methods are general likely to be combined in practice.

4.1 Aggregation of data

In principle, a large number of prominent clustering methods can be used to aggregate geodata. Clustering can be carried out according to spatial, but also to non-spatial criteria such as age or gender. Spatial clusters can be formed, for example, w.r.t. administrative units or by grid lines (Broen et al. 2021). Very promising are grid masking and areal aggregation. The latter is the most frequently used approach for masking georeferenced data, e.g. see Hampton et al. (2010). For instance, regarding so-called disease mapping or in the presentation of morbid-mortality atlases (Ocaña-Riola 2010).

Applying *grid masking*, all original data points are aligned to uniform grid cells (Seidl et al. 2015). The decision maker then selects an appropriate size of the cells. A regularly selected grid cell has a side length of 100m, 125m or 250m, see e.g. Seidl et al. (2015) or Houfaf-Khoufaf and Touya (2021). Applying *grid line masking*, the original points are moved to the nearest edge of the grid cell surrounding them. An actual aggregation occurs when several points have the same nearest grid line and are combined into one single point. Applying *grid centroid masking*, all points in the interior of some cell are moved to the corresponding cell centroid.

Instead of a large number of data points being reflected by one single point, the surrounding area can represent a cloud of data points. An interesting contribution treating Adaptive Areal Elimination (AAE) and a continuation of the Adaptive Areal Masking (AAM) is found in Kounadi and Leitner (2016). Although aggregation methods generally promise a high level of data protection, the original data is heavily altered, resulting in a loss of spatial information. Another major task is the modifiable areal unit problem (MAUP), describing the fact that aggregated data values always depend on the specific setting of boundaries. As we see

in the example in Figure 1, boundaries are decisive in the evaluation of aggregation values.

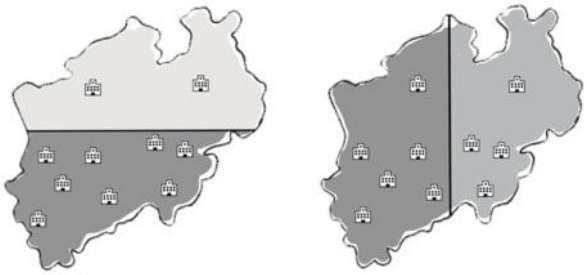


Figure 1 : MAUP using the example of hospital density

In figure 1 on the left, there is a low (light gray) and high density (dark gray) of hospitals with horizontal boundaries. If the boundary is drawn vertically (right example), there is a medium density (medium gray) in addition to a high density. In terms of hospital construction, less need for action could be interpreted in the case on the right if the original data is not available.

4.2 Perturbing of data

Affine point transformations are the most basic geographic masking methods. Affine point transformation shows the advantage that the relative position of points is maintained, i.e. it preserves spatial patterns. At the same time, a data attacker can obviously succeed in re-identifying a large part, if not the entire dataset, based on just two unmasked points. In general, it can be observed that this approach has not gained acceptance in the scientific community. For this reason, these methods will not be discussed in detail.

The following data perturbing methods are considered to be suitable for the anonymization of health data. It is possible and even recommended to use these methods adaptively depending on the population density. Voronoi diagrams have been used in mathematics since the beginning of the 20th century. Seidl et al. (2015) were the first to apply the concept of *Voronoi masking* in the context of geodata. According to them, Voronoi polygons are placed around the points to be masked. Each point defines the center of gravity of its surrounding polygon. In addition, the polygons have the property that their boundary lines always run exactly in the middle of two points. After the polygons have been created, each point is placed on the nearest edge of its surrounding polygon. In contrast to the aggregative methods, in case of Voronoi msking it is generally impossible for the data user to recognize the margins of the aggregates.

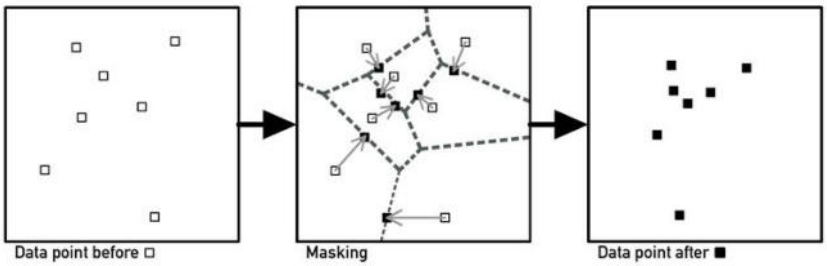


Figure 2 : Voronoi masking

The advantage is that the displacement distance in densely populated areas is smaller than in sparsely populated ones. Thus, the pattern of anonymized points closely resembles the original distribution. Another advantage is that no relocated point is placed on another household, while it can be critically questioned whether Voronoi masking can be meaningful at all in a complete dataset, as the spatial *k*-anonymity could in some areas reduce to $k = 2$ (compare figure 2).

The *donut masking* method is a continuation of the random distribution within a circular region (Hampton et al. 2010, Cremer et al. 2024). In addition to a circular region with the radius r_{max}, into which some point is randomly displaced, an inner ring with the radius r_{min} is defined around the original location, beyond which the displacement must extend. Thus, two radii, which resemble the shape of a donut, determine the minimum and maximum distance between which the new coordinates of the point have to lie. If, in addition to the torus shape, there are further restrictions with regard to the new placement, so-called "eaten donuts" may result. Examples of such restrictions are borders of any kind or uninhabitable places such as parks, bodies of water or mountains. Regarding adaptive donut masking, both the minimum and maximum distance are selected depending on the population density at the original location. If there is a high population density, radii can be selected smaller than in rural areas.

The application of a suitable *Bimodal Gaussian* distribution is in some ways related to donut masking. Instead of the uniformly distributed probability, where the masked point is placed between some inner and the outer radius, a bimodal Gaussian distribution is used for the random distance function (Houfaf-Khoufaf and Touya 2021). The coordinates of the original location $\left(x_i^{orig}, y_i^{orig}\right)$ form the coordinate origin. A distance d_i to this origin is chosen at random from a normal distribution with expectation μ and standard deviation σ, that is, $X \sim N\ (\mu; \sigma)$, resulting in masked coordinates $\left(x_i^{\text{mask}}, y_i^{\text{mask}}\right)$. Generally, different values of μ and σ are tested and several calculation runs are carried out so that an optimum result can be achieved. The displacement angle θ_i is also selected randomly, namely from the uniform distribution $X \sim G\ (0; 2\pi)$. Hence, it follows:

$$\left(x_i^{\text{mask}}, y_i^{\text{mask}}\right) = \left(x_i^{orig} + d_i * \cos(\theta_i)\,, y_i^{orig} + d_i * \sin(\theta_i)\right)$$

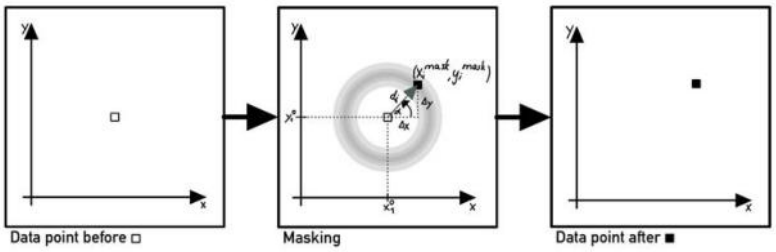


Figure 3 : Bimodal Gaussian masking

Within the darker inner area of the torus (middle of figure 3) the probability of placement is highest. The probability of placement decreases towards the outer and inner edges of the ring before it converges to zero (white area).

The *location swapping* method introduced by Zhang et al. (2017) begins with the creation of a buffer zone with a defined radius around the point to be moved. This radius is defined adaptively depending on the population density around. An address is then randomly selected from the available (real existing) addresses in that zone. The point is then shifted to this address. The strength of data swapping obviously lies in maintaining the geographical distribution of points, while multivariate correlations can be affected depending on specific users' interests.

5 Conclusion and future work

Anonymized data must meet both, protection and quality requirements. In order to compare elaborated methods, optimal parameterizations for achieving specified data protection levels are approximated in future project work, while preserving the associated analytical validity of

the data best possible. Aiming to establish a ranking of the pre-selected methods and to provide suitable combinations of them for different datasets and ways of data access. The overarching goal is to derive recommendations for application and tools to generate anonymized research data of highest possible analysis potential. A scientifically sound portfolio of methods must be developed in a discursive process involving all stakeholders in order to meet the strict data protection requirements on the one hand and the diverse user requirements on the other. Here, a distinction should be made between the possible ways of accessing data, including data dissemination as standardized scientific use files, as what are called campus-files for teaching purposes, at a guest scientist's workstation or by remote data access (Lenz 2011).

Germany has transposed the EU INSPIRE directive into national law with the Geodata Utilization Ordinance and has committed itself to making those INSPIRE-relevant data available as open data. The draft Health Data Usage Act already aims to establish a decentralized health data infrastructure with a central data usage and coordination office. In particular, this should facilitate the usability of the data for public welfare purposes. The current research project *AnigeD*, funded by the German Ministry of Research and Education, is intended to lay the methodological foundation for this. It is therefore important to carry out empirical studies with real world geodata to further compare the methods presented here.

Bibliographie

Armstrong, M.P., Rushton, G. and Zimmermann, D.L. (1999), Geographically Masking Health Data to preserve Confidentiality, *Statistics in Medicine,* 18, pp. 497-525

Broen, K., Rob, T. and Jon, Z. (2021), Measuring the impact of spatial perturbations on the relationship between data privacy and validity of descriptive Statistics, *International Journal of Health Geographics,* 20 (3)

Cremer, S., Jehmlich, L. and Lenz, R. (2024), Masking georeferenced health data - an analysis taking the example of partially synthetic data on sleep disorder, *Privacy in Statistical Databases, Domingo-Ferrer and M. Önen (Eds.),* LNCS 14915, pp. 297–309

Dwork, C. (2006), Differential privacy, *International colloquium on automata, languages, and programming,* pp. 1-12

Eurostat (2023). Statistical confidentiality and personal data protection, see: https://ec.europa.eu/eurostat/de/web/microdata/statistical-confidentiality-and-personal-data-protection

Hampton, K.H., Fitch, M.K., Allshouse, W.B., Doherty, I.A., Gesink, D.C., Leone, P.A. and Miller, W.C. (2010). Mapping Health Data: Improved Privacy Protection With Donut Method Geomasking, *American Journal of Epidemiology,* 172 (9), pp. 1062–1069

Houfaf-Khoufaf, W. and Touya, G. (2021). Geographically Masking Addresses to Study COVID-19 Clusters, *Univ Gustave Eiffel*

Koller, D., Wohlrab, D., Sedlmeir, G. and Augustin, J. (2020), Geografische Ansätze in der Gesundheitsberichterstattung, *Bundesgesundheitsblatt,* 63, pp. 1108–1117

Kounadi, O. and Leitner, M. (2016), Adaptive areal elimination: A transparent way of disclosing protected spatial datasets, *Computers, Environment and Urban Systems,* 57, pp. 59–67

Kwan, M., Casas, I. and Schmitz, B. (2004), Protection of Geoprivacy and Accuracy of Spatial Information: How Effective Are Geographical Masks? *Cartographica,* 39(2), pp. 15-28

Lenz, R. (2011), On the way to remote access to German official microdata, Statistique et nouvelles technologies de l'information, *Revue des Nouvelles Technologies de l'information*, pp. 125-138 (paper presented at the 41ème Journées de Statistique de la SFdS, Bordeaux 2009)

Ocaña-Riola, R. (2010), Common errors in disease mapping, *Geospatial Health,* 4 (2), pp. 139-154

Seidl, D.E., Paulus, G., Jankowski, P., and Regenfelder, M. (2015), Spatial obfuscation methods for privacy protection of household-level data, *Applied Geography,* 63, pp. 253-263

Zhang, S., Freundschuh, S.M., Lenzer, K., Zandbergen, P.A. (2017), The Location Swapping Method for Geomasking, *Cartography and Geographic Information Science,*44 (1), pp. 22-34

Inférence causale 1

AdaptiveCausalWVAE: Certifiably Robust and Stable Causal Disentanglement

Manal Benhamza [1] & Marianne Clausel [2] & Myriam Tami[3]

[1] *Paris-Saclay University, CentraleSupélec, MICS Lab, France, manal.benhamza@centralesupelec.fr*
[2] *Lorraine University, Elie Cartan Institute, IECL, France, marianne.clausel@univ-lorraine.fr*
[3] *Paris-Saclay University, CentraleSupélec, MICS Lab, France, myriam.tami@centralesupelec.fr*

Résumé. Les Variational Autoencodeurs (VAE) fournissent un cadre permettant l'apprentissage de représentations latentes à partir de données de grande dimension. L'hypothèse d'indépendance des VAEs engendre la construction de facteurs latents indépendants que nous proposons d'interpréter comme des facteurs latents sous-jacents du phénomène modélisé par le VAE. Or dans de nombreux cas applicatifs, l'hypothèse d'indépendance des variables latentes (VLs) sous-jacente à ces modèles se révèle inadéquate, car elle ignore les relations causales inhérentes aux phénomènes modélisés. Par conséquent, les techniques d'apprentissage des représentations gagneraient en pertinence en intégrant explicitement la possibilité de modéliser aussi ces relations causales et ainsi apprendre une représentation latente structurée et causale. Les approches actuelles, telles que CausalVAE et DEAR, adressent en partie les structures causales, mais demeurent limitées quant aux contraintes architecturales et la nature des relations permises entre les VLs (uniquement linéaires). Nous proposons ainsi un nouveau cadre, appelé " AdaptiveCausalWVAE ", reposant sur un réseau neuronal contraint par la continuité de Lipschitz et des fonctions d'activation spline, permettant d'apprendre des représentations causalement structurées avec des relations non linéaires entre les VLs et donc plus complexes. De surcroît, notre modèle exploite les propriétés de la distance de Wasserstein pour garantir un entraînement stable. Les expériences menées sur des jeux de données synthétiques et réels montrent qu'AdaptiveCausalWVAE atteint efficacement un meilleur démélage des VLs tout en apprenant des relations non linéaires entre VLs, et se révèle plus stable et robuste. De plus, nous analysons les marges de robustesse du modèle en exploitant la continuité de Lipschitz.

Mots-clés. Apprentissage de représentations causales, Fonctions spline, Robustesse Lipschitz.

Abstract. Disentangled representation learning seeks to extract meaningful, underlying latent factors from high-dimensional data, often using frameworks like Variational Autoencoders (VAE). However, the latter's assumption of independent latent variables (LVs) is unrealistic due to real-world causal relationships. The disentangling techniques would hence represent the underlying modeled phenomenon more if they also reflect the causal relationships between the LVs . Existing methods, such as CausalVAE and DEAR, address causal structure learning with linear relationships between LVs but struggle with complex relationships and architectural limitations. We, therefore, propose a novel framework AdaptiveCausalWVAE, which employs a Lipschitz-constrained neural network with spline activation functions to learn latent causal disentangled representations. Our model also leverages the

properties of Wasserstein Divergence for stable training. Experiments conducted on synthetic and real datasets demonstrate that AdaptiveCausalWVAE effectively achieves better causal disentanglement, is more stable and robust. Additionally, we analyze the model's robustness margins by using Lipschitz continuity.

Keywords. Causal disentanglement learning, Spline functions, Lipschitz robustness.

1 Introduction

Disentangled representation learning is essential for capturing semantically meaningful representations of observations, operating under the hypothesis that high-dimensional data can be generated by simpler, meaningful factors known as ground-truth factors. Frameworks like Variational Autoencoders (VAEs) facilitate this by using an encoder to map input data into a continuous latent space and a decoder for the reconstruction task. Their stochastic nature introduces variability in the learned representations, enabling the generation of samples resembling the input data. However, learning disentangled representations with VAEs enforces independence among the LVs. Yet, the independence assumption may not always hold true in real-world scenarios where factors do likely have causal relationships. As a consequence, existing disentangling models do struggle with causally related factors (Shen et al. 2022). CausalVAE (Yang et al. 2021) injected a Causal layer to the traditional VAE to transform the independent LVs into a causal representation. Nevertheless, the implemented linear causal equation in the causal layer hinders capturing complex (i.e, non linear) relationships between the data. While DEAR (Shen et al. 2022) uses a Structural Causal Model (SCM) to define the prior distribution, and employs Generative Adversarial Networks (GAN) for training, it is also limited because GANs lack the encoder part and hence do not enable learning intermediate input encodings. To overcome these limitations, we therefore propose a new framework called " AdaptiveCausalWVAE " that exploits a Lipschitz-constrained neural network with spline activation functions to enable more complex causal representation learning while controlling robustness. For the training phase, we propose a new loss function that includes Wasserstein's evidence lower bound loss for leveraging its powerful properties such as more stability, symmetry, the ability to take into account spatial relations between distributions and stay well defined even if distribution support overlap happens. In the same line as (Yang et al. 2021) we use additional label information to learn the causal structure of disentangled latent factors. The contributions of our work are precisely the following:

1. We introduce a 1 Lipschitz neural network with spline functions that permit to learn non-linear causal relations between disentangled latent factors.
2. We train the encoder with the Wasserstein Distance to take advantage of its symmetry and stability properties.
3. Based on experiments, we show that AdaptiveCausalWVAE can achieve causal disentanglement and perform interventions on both synthetic and real datasets.
4. We analyze the robustness margins of a general CausalVAE leveraging the Lipschitz constants of its components.

2 CausalVAE

The CausalVAE differs from the traditional VAE by its causal layer, which is designed to learn the causal relationships between the LVs. The model is composed of several key components. First, the encoder takes as input the data $x \in \mathcal{X}$, its labels $u \in \mathcal{U}$ and is responsible for outputting a normal Gaussian distribution over the latent space. The labels u help to address the identifiability issues of the model and we denote $\mathbf{S}$ samples from the latent space. Next, the Causal Layer applies the linear SCM in Equation 1, to transform the independent variables $\mathbf{S}$ into endogenous causal variables $\mathbf{Z}$.

$$\mathbf{Z} = \mathbf{A}^\top \mathbf{Z} + \mathbf{S} = (\mathbf{I} - \mathbf{A}^\top)^{-1}\mathbf{S} \tag{1}$$

where $\mathbf{A}$ is the adjacency matrix of the Directed Acyclic Graph (DAG). Considering that the latter adjacency matrix $\mathbf{A}$ is expressed as $\mathbf{A} = [\mathbf{A}_1 \mid \ldots \mid \mathbf{A}_n]$, where each $\mathbf{A}_j$ is the weight vector that encodes the causal force from variable $\mathbf{Z}_i$ to variable $\mathbf{Z}_j$, the mask layer propagates the effect of the causal parents to their children using Equation 2.

$$\mathbf{Z}_j = g_j(\mathbf{A}_j \circ \mathbf{Z}; \eta_j) + \epsilon_j \tag{2}$$

where g_j represents a set of mild non-linear invertible functions, η_j are their respective parameters, and ϵ_j are noise variables. Finally, the decoder acts as the pipeline generator, restoring the data from the causal representation. In addition to its ability to reconstruct data, the CausalVAE also facilitates the generation of counterfactual samples by intervening on variables in the causal representation $\mathbf{Z}$. However, its performance in terms of reconstruction and intervention is suboptimal on datasets where the relationships between latent causal variables are non-linear, such as on CelebA (Liu et al. 2015). This limitation arises from the use of a linear causal equation in the causal layer. Moreover, it is important to note that the training process of this model can be unstable due to the Kullback-Leibler Divergence metric, which is neither symmetric nor robust to variations in dataset size. We hence, propose in Section 3 a method to overcome CausalVAE limitations, mainly the used linear causal equation and the unstable learning dynamics. In Section 4, we analyze the robustness margins of our proposed model showing that the 1 Lipschitz constraint helped gaining in robustness.

3 Proposed model: AdaptiveCausalWVAE

The proposed model's architecture is illustrated in Figure 1. We added, in the causal layer, a neural network where the layers are enforced to have a Lipschitz constant strictly inferior to 1, and the spline functions are used as trainable adaptive activation functions, in order to learn the non linear causal relationships between the concepts in data. We also wanted to ensure a stable training of our model, and therefore replaced the Kullback Leibler divergence by the Wasserstein distance. Our choice for the divergence metric is motivated by the fact that, unlike the Kullback-Leibler (KL) divergence which is asymmetric and can become undefined when the distributions have no overlapping support, the Wasserstein distance is symmetric and remains well-defined regardless of whether the distributions overlap or not. As for the spline functions (Aziznejad et al. 2020), the latter have gained attention as activation

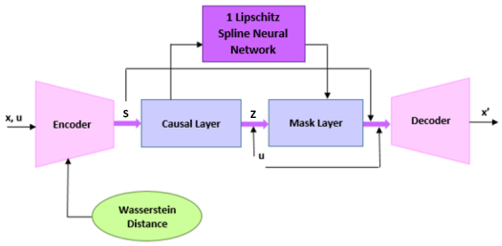


Figure 1: AdaptiveCausalWVAE diagram

functions in neural networks due to their unique properties of smoothness and flexibility. Furthermore, unlike traditional activation functions such as ReLU or Sigmoid (Dubey et al. 2022), spline functions offer the advantage of piecewise polynomial approximation, which can fit complex, non-linear patterns in the data. Combining 1 Lipschitz constrained neural networks with spline activation functions is justified in (Neumayer et al. 2022), where it has been proved that a constrained neural network with spline activations can approximate a wide range of functions. In our case, we specifically chose 1-Lipschitz neural splines to control the Lipschitz constants of the causal layers and adapt them to our robustness needs.

4 Robustness Evaluation

Building on the definition of the robustness of a VAE proposed by (Camuto et al. 2020), we show that the Lipschitz continuity can be further exploited to calculate the robustness margins of both the linear CausalVAE and AdaptiveCausalWVAE. The latter do belong to a family that we name General CausalVAE and where the causal layer implements the general SCM in Equation 3, with f and h being element-wise transformations operating on each individual element of their input vectors.

$$\mathbf{Z} = f\left((\mathbf{I} - \mathbf{A}^\top)^{-1} h(\mathbf{S})\right) \tag{3}$$

For the linear CausalVAE, f and h correspond to the identity function, whereas for the AdaptiveCausalWVAE f is a 1 Lipschitz constrained neural network and h is the identity function. Using the above form of a SCM, we express in Theorem 1 the probability bound of a model from the general CausalVAE family, following what has been done in (Camuto et al. 2020) for the traditional VAE.

Theorem 1 (General CausalVAE Probability Bound) *Let* $\mathbf{Z}_\delta = f((\mathbf{I}-\mathbf{A}^T)^{-1}h(\mathbf{S}_\delta)),\mathbf{z}_{\neg\delta} = f((\mathbf{I}-\mathbf{A}^T)^{-1}h(\mathbf{S}_{\neg\delta}))$, $q_\phi(\mathbf{S}|x,u) = \mathcal{N}(\mathbf{S};\mu_\phi(x,u),\mathrm{diag}(\sigma^2_\phi(x,u)))$ *be the used parametrization of the encoder and* d_z *the dimension of the latent space. We assume that the deterministic component of the general CausalVAE decoder* $g_\theta(\cdot)$ *is* a*-Lipschitz, the encoder mean* $\mu_\phi(\cdot)$ *is* b*-Lipschitz, and the encoder standard deviation* $\sigma_\phi(\cdot)$ *is* c*-Lipschitz,* f *and* h *the element-wise transformations of the general non-linear SCM are also respectively* d *and* e *Lipschitz. Then for any* $r \in \mathbb{R}_+$*, any* $x \in \mathcal{X}$ *and its corresponding label* $u \in \mathcal{U}$*, and any input perturbation* $\delta \in \mathcal{X}$*,*

$$P\left[\|g_\theta(\mathbf{Z}_\delta) - g_\theta(\mathbf{Z}_{\neg\delta})\|_2 \leq r\right] \geq 1 - \min\{p_1(x,u), p_2(x,u)\} \tag{4}$$

where

$$p_1(x,u) := \min\left(1, \frac{k^2(b^2\| \ \delta\|_2^2 + (c\| \ \delta\|_2 + 2\|\sigma_\phi(x,u)\|_2)^2)}{r^2}\right),$$

$$p_2(x,u) := \begin{cases} C(d_z)\frac{n(x,u)^{d_z/2}\exp\left(-\frac{n(x,u)}{2}\right)}{(n(x,u)-d_z+2)} & \text{if } \frac{r}{k} - b\|\delta\|_2 \geq 0; d_z \geq 2; n(x,u) > d_z - 2 \\ 1 & \text{otherwise} \end{cases}$$

$$\begin{cases} n(x,u) := \left(\frac{\frac{r}{k}-b\|\delta\|_2}{c\|\delta\|_2+2\|\sigma_\phi(x,u))\|_2}\right)^2 \\ C(d_z) := \frac{1}{\sqrt{\pi}}\exp\left(\frac{1}{2}\left(d_z - (d_z-1)\log d_z\right)\right) \\ k = aed\|(\mathbf{I} - \mathbf{A}^T)^{-1}\|_2 \end{cases}$$

also

$$\mathbf{S}_\delta \sim q_\phi(\mathbf{S}|x+\delta,u) \text{ and } \mathbf{S}_{\neg\delta} \sim q_\phi(\epsilon|x,u).$$

We can observe that the obtained probability bounds are the same as for a traditional VAE with the difference of considering the Lipschitz constants of the causal layer, which is assumed to be Lipschitz continuous. The local robustness margins corresponding to the probability bounds in Equation 4 are also expressed in terms of the Lipschitz constants of all components of a general CausalVAE model in Lemma 1.

Lemma 1 (Local Margin Bound General CausalVAE) *Given the assumptions of Theorem 1 and some $\epsilon \in [0,1)$, the (r,ϵ)-robustness margin of the general CausalVAE on input x with label u is the following,*

$$R_{(r,\epsilon)}(x,u) \geq \max\{m_1(x,u), m_2(x,u)\}$$

for

$$m_1(x,u) = \frac{-4c||\sigma_\phi(x,u)||_2 + \sqrt{\Delta}}{(c^2+b^2)}.$$

where

$$\Delta = (4c||\sigma_\phi(x,u)||_2)^2 - 4(c^2+b^2)\left(4||\sigma_\phi(x,u)||_2 - (1-\epsilon)\left(\frac{r}{k}\right)^2\right)$$

and

$$m_2(x,u) := \sup\{||\delta||_2 : p_2(\delta,x,u) \leq (1-\epsilon)\}$$

where $p_2(\delta,x,u)$ is expressed in Theorem 1.

The global margin bounds can be obtained by fixing the encoder standard deviation as a hyperparameter of the model and hence setting $\sigma_\phi(x,u) = \sigma$. The intuition behind the local and global margin bounds will be explained using applied examples on the Pendulum and CelebA datasets. One of the advantages of the expressed robustness margin bounds for a general CausalVAE is that they enable the comparison of the robustness of the CausalVAE and AdaptiveCausalWVAE models solely based on their respective coefficient k, and more specifically from the product of the Lipschitz constants of their respective causal layers *ed*. The AdaptiveCausalWVAE verifies $e = 1, d < 1$ whereas for the linear CausalVAE $e = 1, d = 1$ and hence $k_{\text{AdaptiveCausalWVAE}} < k_{\text{CausalVAE}}$, implying, based on experiments, that the AdaptiveCausalWVAE has larger robustness margin bounds and is hence more robust.

5 Results

The proposed architecture has enabled us to learn non-linear relations between the causal variables therefore boosting the Mutual and Total Information Coefficients on both CelebA and Pendulum, as shown in Table 1. Plots on the stability and robustness for both datasets will be presented during the conference.

Model	Pendelum		CelebA	
	MIC (%)	TIC (%)	MIC (%)	TIC (%)
CausalVAE	95.1 ± 2.5	81.6 ± 6.2	92.5 ± 2.9	80.4 ± 7.9
AdaptiveCausalWVAE	**96.5 ± 1.5**	**83.5 ± 5.4**	**94.8 ± 0.9**	**83 ± 8.4**
Condition-VAE (Kihyuk S. 2015)	93.8 ± 1.9	80.5 ± 6.1	78.8 ± 3.7	66.1 ± 2.2

Table 1: Performance Evaluation

6 Acknowledgements

This research was funded in whole or in part by the National Research Agency (ANR) under the project "ANR-23-CE23-0008-01".

References

Aziznejad, S., Gupta, H., Campos, J., and Unser, M. (2020). "Deep Neural Networks With Trainable Activations and Controlled Lipschitz Constant". In: *IEEE Transactions on Signal Processing*.

Camuto, A., Willetts, M., Roberts, S., Holmes, C., and Rainforth, T. (2020). "Towards a Theoretical Understanding of The Robustness of Variational Autoencoders". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*.

Dubey, S-R., Singh, S-K., and Chaudhuri, B-B. (2022). "Activation functions in deep learning: A comprehensive survey and benchmark". In: *Neurocomputing* 503, pp. 92–108.

Kihyuk S. Honglak L., Xinchen Y. (2015). "Learning Structured Output Representation using Deep Conditional Generative Models". In: *NeurIPS*.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). "Deep Learning Face Attributes in the Wild". In: *2015 IEEE International Conference on Computer Vision (ICCV)*.

Neumayer, S., Goujon, A., Bohra, P., and Unser, M. (2022). *Approximation of Lipschitz Functions using Deep Spline Neural Networks*. eprint: 2204.06233 (cs.LG).

Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z., and Zhang, T. (2022). "Weakly Supervised Disentangled Generative Causal Representation Learning". In: *Journal of Machine Learning Research* 23.241, pp. 1–55.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. (2021). "Causalvae: Disentangled representation learning via neural structural causal models". In: *CVPR*, pp. 9593–9602.

Inférence de graphe de différence par découverte causale non-paramétrique

Daria Bystrova [1] & Emilie Devijver [1]

[1] *Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France*

Résumé. Comparer les relations causales entre différentes populations est essentiel dans de nombreux domaines. Ce travail traite de l'inférence d'un réseau de différences via la découverte causale en proposant une nouvelle méthode qui fonctionne sans hypothèses semi-paramétriques et qui prend en compte des données continues, discrètes et mixtes. Nous supposons que les deux populations diffèrent en raison d'interventions sur un modèle causal structurel inconnu, et nous introduisons le concept de fidélité différentielle. Notre approche inclut des garanties théoriques et démontre de bonnes performances sur des ensembles de données simulées. Enfin, nous appliquons notre méthode à un jeu de données réel afin d'illustrer son utilité pratique.

Mots-clés. Inférence causale, graphe de différence, algorithme PC

Abstract. Comparing causal relationships across populations is essential in various fields. This work addresses the inference of a difference network via causal discovery by proposing a novel method that operates without semi-parametric assumptions and handles continuous, discrete, and mixed data. We assume that the two different population arise from interventions on an unknown structural causal model, and we introduce the concept of differential faithfulness. Our approach includes theoretical guarantees and demonstrates good performance on simulated datasets, and we apply our approach to a real-world dataset to illustrate its practical utility.

Keywords. Causal inference, difference graph, PC algorithm

1 Introduction et état de l'art

Identifier les structures causales et comparer ces relations entre différentes populations permet de détecter des variations significatives. Par exemple, en médecine, cela permet d'adapter les traitements aux sous-groupes de patients, tandis qu'en écologie, cela révèle des changements dans la résilience des écosystèmes. Les méthodes classiques d'inférence causale reposent souvent sur l'identification séparée des graphes causaux pour chaque sous-population, puis leur comparaison. Toutefois, cette approche est sous-optimale, augmentant les risques d'erreurs statistiques et omettant des nuances importantes, comme les variations d'intensité des relations causales. Pour surmonter ces limitations, des approches récentes se concentrent sur l'inférence directe de graphes de différences, où un arc représente une variation dans les liens causaux entre les populations.

Les travaux existants, tels que ceux utilisant des modèles graphiques gaussiens ou des modèles de Markov, comparent directement des structures de dépendance conditionnelle. Certaines méthodes pour les graphes causaux orientés (DAGs), comme l'algorithme DCI [Wang et al., 2018] et iSCAN [Chen et al., 2024], utilisent des tests d'indépendance ou une approche semi-paramétrique, mais reposent sur des hypothèses fortes telles que la linéarité ou la normalité des distributions, et sont principalement adaptées aux variables continues.

Dans cet article, nous introduisons Implicit Difference Inference (IDI), un nouvel algorithme qui infère des graphes de différences en une seule étape, sans nécessiter d'hypothèses paramétriques sur la distribution des données. Contrairement aux méthodes traditionnelles basées sur la fidélité causale, IDI repose sur la fidélité différentielle, une condition plus souple qui renforce la robustesse de l'inférence causale dans des environnements complexes. Ce cadre est applicable à des données continues, discrètes et mixtes, ce qui étend son domaine d'application.

Nos contributions principales sont les suivantes : 1/ Nous proposons la fidélité différentielle, une alternative flexible à l'hypothèse de fidélité classique pour l'inférence de graphes de différences ; 2/ Nous développons IDI, un algorithme permettant une inférence efficace des graphes de différences avec des hypothèses minimales. Bien que ce résumé se concentre sur la théorie, des résultats expérimentaux seront également présentés.

2 Préliminaires

Nous considérons un *modèle causal structurel* (SCM) [Pearl, 2000], noté $\mathcal{M}^*$, composé d'un ensemble de variables cachées $\mathbb{U}$, d'un ensemble de variables observées $\mathbb{V} = \{X_1, \ldots, X_p\}$, et de mécanismes causaux $\mathbb{F}$ reliant ces variables. Ce modèle induit un graphe orienté acyclique (DAG) $\mathcal{G}^* = (\mathbb{V}, \mathbb{E}^*)$, où une arête $X_i \to X_j$ représente une relation causale. Nous supposons la *suffisance causale*, c'est-à-dire que toutes les causes communes des variables observées sont incluses. Une *intervention* modifie un mécanisme causal et peut altérer la distribution conditionnelle associée à une variable. Un *environnement* e correspond à un SCM résultant d'interventions appliquées à $\mathcal{M}^*$, engendrant un DAG $\mathcal{G}^e$ et une distribution conjointe $\mathbb{P}^e$. Les interventions peuvent être *structurelles* (rupture totale des relations parentales) ou *douces* (modification partielle des liens). Suivant Pearl [2000], nous supposons la modularité des mécanismes causaux : une intervention sur X_j affecte uniquement $\mathbb{P}^e(X_j|\mathrm{Pa}_{\mathcal{G}^*}(X_j))$, laissant les autres distributions conditionnelles intactes. Nous utilisons également les concepts standard de *d-séparation* pour caractériser les dépendances conditionnelles dans les DAGs.

Dans le cas de la différence entre deux DAGs, nous considérons deux environnements e^1 et e^2 résultant d'interventions sur $\mathcal{M}^*$, chacun induisant un DAG $\mathcal{G}^1$ et $\mathcal{G}^2$. Nous supposons que ces DAGs partagent un ordre topologique commun, empêchant les inversions d'arêtes [Malik et al., 2024]. Pour analyser les différences causales entre ces environnements, nous définissons le *DAG différentiel*, qui capture les changements structurels entre $\mathcal{G}^1$ et $\mathcal{G}^2$:

Définition 1 (DAG différentiel)**.** *Un graphe $\mathcal{D} = (\mathbb{V}, \mathbb{E})$ est un* DAG différentiel *entre deux SCMs $\mathcal{M}^1$ et $\mathcal{M}^2$ si chaque arête $X_i \to X_j \in \mathbb{E}$ indique que X_i est parent de X_j dans $\mathcal{G}^*$ et que l'effet direct de X_i sur X_j diffère entre $\mathcal{M}^1$ et $\mathcal{M}^2$.*

3 La méthode IDI

L'inférence du graphe des différences vise à identifier les relations causales inchangées et celles qui varient entre deux environnements. Dans cette section nous introduisons un outil graphique pour travailler sur ce graphe, et une méthode qui découvre ce graphe à partir de données observationnelles.

Diff-séparation Nous introduisons la notion de diff-séparation pour établir un lien entre l'information mutuelle conditionnelle et les mécanismes causaux. Contrairement aux DAGs classiques, où tous les chemins peuvent être pertinents, nous nous concentrons spécifiquement sur les chemins qui incluent des nœuds affectés par des interventions.

Définition 2. *Soit $\mathcal{M}^*$ un SCM. Considérons deux environnements e^1 et e^2, où les SCM respectifs $\mathcal{M}^1$ et $\mathcal{M}^2$ sont obtenus en appliquant des interventions à $\mathcal{M}^*$. Nous désignons par $\mathcal{G}^1$ et $\mathcal{G}^2$ les graphes causaux correspondants, et par $\mathbb{I}^1$ et $\mathbb{I}^2$ les ensembles de variables ayant subi des interventions dans chaque environnement. Un ensemble de variables $\mathbb{Z}$ est dit* diff-séparer *une paire ordonnée de nœuds (X_i, X_j) dans $(\mathcal{G}^1, \mathcal{G}^2)$ si :*

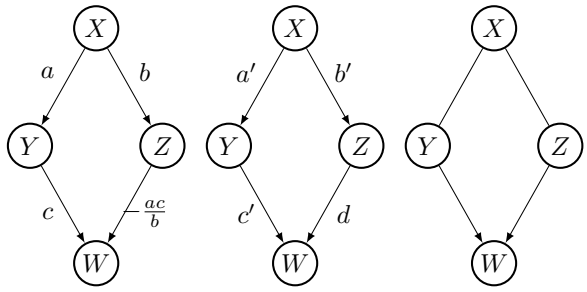
- *Dans $\mathcal{G}^1$, $\mathbb{Z}$ bloque tous les chemins actifs entre X_i et X_j qui passent par au moins un nœud $X_k \in \mathbb{I}^1$ ayant un arc entrant sur ce chemin.*
- *Dans $\mathcal{G}^2$, $\mathbb{Z}$ bloque tous les chemins actifs entre X_i et X_j qui passent par au moins un nœud $X_l \in \mathbb{I}^2$ ayant un arc entrant sur ce chemin.*

Le lemme suivant établit un lien direct entre la diff-séparation et la structure du graphe des différences.

Lemme 1. *Soit $\mathcal{M}^*$ un SCM. Considérons deux environnements e^1 et e^2, où les SCM respectifs $\mathcal{M}^1$ et $\mathcal{M}^2$ sont obtenus en appliquant des interventions à $\mathcal{M}^*$. Nous désignons par $\mathcal{G}^1$ et $\mathcal{G}^2$ les graphes causaux correspondants, et par $\mathcal{D} = (\mathbb{V}, \mathbb{E})$ le graphe des différences entre $\mathcal{M}^1$ et $\mathcal{M}^2$. Pour toute paire de nœuds $(X_i, X_j) \in \mathbb{V}^2$, il existe un sous-ensemble $\mathbb{Z} \subseteq \mathbb{V} \setminus \{X_i, X_j\}$ qui diff-sépare X_i de X_j dans $(\mathcal{G}^1, \mathcal{G}^2)$ si et seulement si $X_i \notin Pa_{\mathcal{D}}(X_j)$.*

Information mutuelle conditionnelle et hypothèse de fidélité différentielle Pour tester la diff-séparation, nous comparons l'information mutuelle conditionnelle entre les distributions associées à chaque SCM. Cependant, une égalité de cette mesure ne garantit pas toujours des mécanismes causaux identiques, ce qui motive l'introduction d'une hypothèse pour affiner cette distinction.

Les méthodes classiques basées sur les contraintes, qui visent à découvrir un DAG causal à partir de données observationnelles, reposent généralement sur l'hypothèse de fidélité. En substance, la fidélité garantit que la d-séparation dans le graphe correspond à l'indépendance statistique dans les données. Formellement, sous fidélité, X_i et X_j sont d-séparés par $\mathbb{Z}$ si et seulement si $I(X_i, X_j \mid \mathbb{Z}) = 0$. Cependant, lors de la construction d'un graphe des différences basé sur des contraintes de dépendance, l'hypothèse de fidélité n'est pas requise, comme l'illustre l'Exemple 1.



(a) $\mathcal{G}_1$ (b) $\mathcal{G}_2$ (c) $\hat{\mathcal{D}}$

FIGURE 1 – Violation de la fidélité mais pas de la fidélité différentielle. L'algorithme PC ne peut pas découvrir $\mathcal{G}_1$, mais IDI peut inférer $\hat{\mathcal{D}}$, un graphe associé à la classe d'équivalence du vrai DAG de différence $\mathcal{D}$.

Exemple 1. *Considérons un SCM linéaire $\mathcal{M}^*$ et deux environnements, définissant respectivement les SCMs $\mathcal{M}^1$ et $\mathcal{M}^2$ avec les DAGs $\mathcal{G}^1$ et $\mathcal{G}^2$ (Fig. 1 (a) et (b)). Les effets causaux directs dans $\mathcal{M}^1$ et $\mathcal{M}^2$ sont représentés par les arcs des graphes.*

Dans $\mathcal{G}^1$, il existe deux chemins distincts entre X et W dont les effets se compensent. Une méthode standard basée sur l'indépendance pourrait à tort conclure que X et W sont indépendants, supprimant ainsi l'arc. Cependant, si les coefficients causaux sont modifiés dans $\mathcal{M}^2$, les dépendances statistiques entre X et W peuvent changer. La méthode IDI prend en compte ces différences et conserve l'arc, sauf si une condition sur les variables intermédiaires Y et Z montre que la dépendance entre X et W reste la même entre les environnements.

Notre méthode évite l'hypothèse de fidélité, la rendant plus robuste aux variations mineures, et ne requiert pas l'hypothèse de minimalité, ce qui la rend applicable à des modèles complexes. Elle repose sur la fidélité différentielle, garantissant la détection des effets causaux entre environnements pour une identification fiable du graphe des différences.

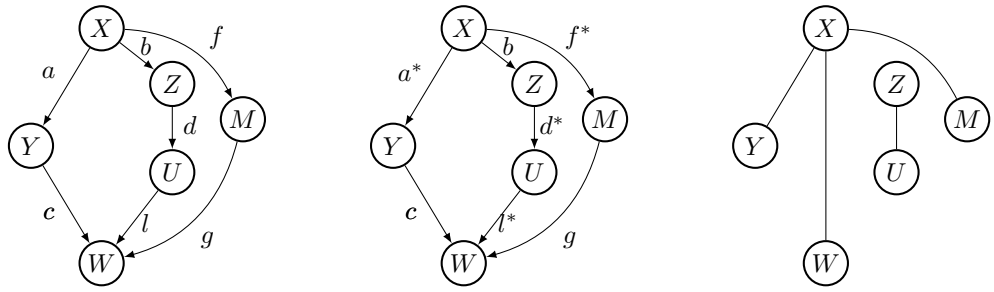
Hypothèse 1 (Fidélité différentielle)**.** *Soit $\mathcal{M}^*$ un SCM. Considérons deux environnements e^1 et e^2, où les SCM respectifs $\mathcal{M}^1$ et $\mathcal{M}^2$ sont obtenus en appliquant des interventions à $\mathcal{M}^*$. Nous désignons par $\mathcal{G}^1$ et $\mathcal{G}^2$ les graphes causaux respectifs et par $\mathbb{P}^1$ et $\mathbb{P}^2$ les distributions associées.*

Nous supposons que

$$I_{\mathbb{P}^1}(X_i, X_j|\mathbb{Z}) = I_{\mathbb{P}^2}(X_i, X_j|\mathbb{Z})$$

si et seulement si $\mathbb{Z}$ diff-sépare X_i de X_j et $\mathbb{Z}$ diff-sépare X_j de X_i dans $(\mathcal{G}^1, \mathcal{G}^2)$.

Cette hypothèse garantit que les interventions causales ont des effets détectables sur les dépendances conditionnelles, permettant ainsi la construction du graphe des différences en capturant avec précision où et comment les dépendances varient entre les environnements. L'hypothèse de fidélité différentielle implique également que les chemins d'annulation identiques (c'est-à-dire les violations de la fidélité) ne peuvent pas persister à travers les deux graphes causaux, et elle restreint l'échange des coefficients causaux entre les graphes.



(a) $\mathcal{G}_1$ (b) $\mathcal{G}_2$ (c) $\hat{\mathcal{D}}$

FIGURE 2 – Example of the error of IDI algorithm with PC adjacency search.

Algorithme d'Inférence de Différence Implicite (IDI) Pour construire le graphe de différence, nous proposons un algorithme inspiré de l'algorithme SGS [Spirtes et al., 2000], conçu à l'origine pour la découverte causale classique. Au lieu de nous appuyer sur la d-séparation, nous utilisons la diff-séparation pour inférer le graphe de différence.

L'algorithme suit trois étapes principales : 1/ Construction du squelette, 2/ Orientation des structures en V, 3/ Règles d'orientation supplémentaires. Le graphe obtenu après les deux premières étapes est appelé le pattern du graphe de différence, noté $pattern(\mathcal{D})$, qui comprend le squelette correct et toutes les structures en V orientées.

Une version modifiée de l'algorithme PC [Spirtes et al., 2000, Colombo and Maathuis, 2014] a été introduite pour accélérer ces premières étapes du SGS en : 1/ considérant des ensembles de conditionnement de taille croissante ; 2/ se concentrant uniquement sur les nœuds adjacents ; 3/ stockant les ensembles de séparation pour une construction efficace du graphe. Cependant, l'algorithme PC ne peut pas être directement adapté à l'inférence de DAG de différence. Un exemple est donné pour illustrer cette limitation.

Exemple 2. *Considérons deux SCMs, $\mathcal{M}^1$ et $\mathcal{M}^2$, définis sur les variables X, Y, Z, W, U, M. Nous supposons des interventions sur les nœuds Y, U, M, modifiant les relations causales directes entre certaines paires : (X, Y), (Z, U), (X, M). L'algorithme commence avec un graphe pleinement connecté, non dirigé, et l'appliqué naïvement échoue dans ce cas, comme l'illustre la figure 2.*

Cet exemple illustre la limite de l'algorithme PC pour inférer un DAG de différence, car sa stratégie de conditionnement basée sur l'adjacence ne permet pas de retrouver le squelette réel. Toutefois, certaines de ses idées sont exploitées, comme le conditionnement séquentiel et l'utilisation des ensembles de séparation pour orienter les structures en V.

La troisième étape de notre méthode consiste en l'orientation supplémentaire des arêtes, à l'aide des règles de Meek, adaptées pour les graphes de différence. Ces règles conduisent à la bonne orientation du graphe de différence, comme le montre le lemme suivant.

Lemme 2. *Considérons un graphe de différence $\mathcal{D} = (\mathbb{V}, \mathbb{E})$ et son pattern $pattern(\mathcal{D})$. L'application successive des règles de Meek dans $pattern(\mathcal{D})$ produit la bonne orientation.*

Le graphe résultant est partiellement orienté, car les tests d'indépendance à eux seuls ne suffisent pas à orienter toutes les paires ou triplets de nœuds. Le théorème suivant montre que l'algorithme IDI construit le vrai graphe de différence partiellement dirigé, avec le squelette correct, les structures en V et les orientations de Meek.

Théorème 1. *Soit $\mathcal{M}$ un SCM. Considérons deux environnements e^1 et e^2, où les SCMs respectifs $\mathcal{M}^1$ et $\mathcal{M}^2$ sont obtenus par interventions sur $\mathcal{M}$. Soit $\mathcal{D} = (\mathbb{V}, \mathbb{E})$ le vrai graphe de différence entre $\mathcal{M}^1$ et $\mathcal{M}^2$, et son pattern pattern($\mathcal{D}$). Sous l'Assumption 1 et en supposant un accès à des informations parfaites d'indépendance conditionnelle pour toutes les paires de variables dans $\mathbb{V}$, l'algorithme IDI produit le vrai DPDAG associé à $\mathcal{D}$.*

4 Conclusion

Nous avons présenté IDI, une méthode d'inférence des DAG de différence pour SCM non paramétriques, capable de gérer des données mixtes. Validée théoriquement, elle offre une approche robuste sans hypothèses strictes. IDI ouvre la voie à des méthodes plus flexibles en science et en médecine, avec des applications en cours sur des jeux de données réels. Ce travail soulève aussi de nouvelles perspectives, comme la caractérisation des classes d'équivalence et l'intégration de variables latentes, renforçant son impact tant théorique que pratique.

References

Tianyu Chen, Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. iscan : Identifying causal mechanism shifts among nonlinear additive noise models. *Advances in Neural Information Processing Systems*, 36, 2024.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(116) :3921–3962, 2014.

Vineet Malik, Kevin Bello, Asish Ghoshal, and Jean Honorio. Identifying Causal Changes Between Linear Structural Equation Models. June 2024. URL `https://openreview.net/forum?id=y31rgYaLaS`.

Judea Pearl. *Causality : Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2000. ISBN 0-521-77362-8.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Yuhao Wang, Chandler Squires, Anastasiya Belyaeva, and Caroline Uhler. Direct estimation of differences in causal graphs. *Advances in neural information processing systems*, 31, 2018.

Ensembles de variables causales pour la généralisation de distribution

Théotime Le Goff [1] & Emilie Devijver [1]

[1] *Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France, prenom.nom@univ-grenoble-alpes.fr*

Résumé. La généralisation de distribution est un enjeu clé lors de l'entraînement d'un modèle d'apprentissage automatique. L'objectif est de prédire avec précision des données non observées pendant l'entraînement, qui peuvent être différentes de celles utilisées pour l'apprentissage. Cependant, cette tâche est complexe, car des relations artificielles peuvent être apprises pendant l'apprentissage sans être réellement utiles pour la prédiction hors distribution. Pour aborder ce problème, nous nous appuyons sur le formalisme causal, qui permet de modéliser certains aspects du changement de distribution. Cette approche repose sur le concept d'intervention. Nous proposons une investigation empirique de la capacité d'un modèle à généraliser en utilisant BART (Bayesian Additive Regression Trees), une méthode non paramétrique de prédiction. L'expérience est réalisée sur des données synthétiques afin de déterminer l'ensemble de variables permettant de prédire efficacement des observations issues de distributions modifiées par une intervention.

Mots-clés. Généralisation de domaine, Intervention, Sélection de variables, Arbres de décisions bayésiens.

Abstract. Distribution generalization is a key challenge when training a machine learning model. The goal is to accurately predict unseen data during training, which may be different from the training data. However, this task is challenging because the model may learn artificial relationships that do not contribute to accurate out-of-distribution prediction. To address this issue, we rely on causal formalism, which allows modeling certain aspects of distribution changes. This approach is based on the concept of intervention. We propose an empirical investigation of a model's ability to generalize by using BART (Bayesian Additive Regression Trees), a non-parametric prediction method. The experiment is conducted on synthetic data to determine which set of variables allows effective prediction of observations drawn from distributions modified by an intervention.

Keywords. Domain generalization, Intervention, Feature selection, Bayesian decisions trees.

1 Introduction

Les méthodes d'apprentissage automatique traditionnelles, en particulier celles dédiées à la prédiction, ont prouvé leur efficacité dans de nombreux domaines. Cependant, elles se

heurtent à une difficulté majeure : puisque l'inférence est faite sur des données d'entraînement, toute divergence entre les distributions des données d'entraînement et de test peut détériorer leurs performances. Pour remédier à ce problème, il est essentiel de définir des limites claires sur les variations admissibles entre ces distributions. En effet, si la distribution des données de test devient totalement aléatoire, la généralisation du modèle devient impossible.

Dans notre analyse, nous caractérisons ce changement de domaine à travers le concept d'intervention [3, 5, 8]. En effet, intervenir sur une ou plusieurs variables entraîne une modification de la distribution jointe. Nous estimons que cette approche offre un cadre pertinent pour définir le changement de distribution. Celui-ci s'opère de manière cohérente, préservant certains liens d'origine entre les variables modifiées par ces interventions [7]. Cela suggère que les variables d'entraînement assurant une bonne prédiction hors distribution sont structurées par des relations causales

Un choix judicieux de variables causales semble être utile pour garantir la robustesse d'un modèle face à des données hors distribution. Inclure toutes les variables causalement liées à la cible dans la distribution d'entraînement peut entraîner une redondance d'informations avec sa couverture de Markov [8], qui est théoriquement suffisante pour prédire correctement la cible. Cependant, cet ensemble est moins robuste face aux interventions. Une intervention sur les enfants de la cible supprime leurs relations ainsi que celles des conjoints, tandis que les parents conservent leurs liens. Toutefois, ces derniers peuvent être trop restrictifs et réduire la précision par rapport à d'autres variables.

Nous évaluons empiriquement l'efficacité de différents ensembles de variables causales pour la prédiction en utilisant des données synthétiques. Nous analysons leur capacité prédictive face à des données interventionnelles à l'aide de BART (Bayesian Additive Regression Trees, [1]), une méthode non paramétrique. Cet article est structuré comme suit : la section 2 introduit les concepts essentiels, la section 3 présente le problème et BART, les résultats expérimentaux sont présentés en section 4, et nous concluons avec les perspectives en section 5.

2 Notions préliminaires

Réseau Bayésien Causal. Un réseau bayésien $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ est un graphe dont l'ensemble des sommets $\mathbf{V} = \{1, \ldots, d\}$ représente les variables $X_1, \ldots, X_d$ et dont l'ensemble des arêtes $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ définit les relations de dépendance entre ces variables. Ce graphe est en général acyclique et dirigé (Directed Acyclic Graph, DAG) [5]. Il permet de modéliser la densité jointe $p(\mathbf{X})$ des $X_1, \ldots, X_d$ via des densités conditionnelles. Cette distribution p se réécrit comme le produit des distributions conditionnelles $p_i(X_i \,|\, \mathrm{pa}(X_i))$, où $\mathrm{pa}(X_i)$ désigne les parents de X_i dans $\mathbf{G}$. Pour que ce réseau bayésien $\mathbf{G}$ soit considéré comme causal, où les arêtes représentent des liens causaux, il faut introduire la notion d'intervention [3, 5, 8].

Modèle d'Équations Structurelles. Un autre cadre pertinent pour représenter les relations causales entre variables est le modèle d'équations structurelles (Structural Equation Model, SEM) [8]. Cet outil permet de formaliser ces liens causaux et de modéliser une distribution, qu'elle soit observationnelle ou résultant d'une ou plusieurs interventions. La

structure consiste en une collection de d assignations de la forme

$$X_i = f_i(\mathrm{pa}(X_i), N_i)$$

où N_i désigne un bruit suivant une certaine distribution, pour tout $i \in \mathbf{V}$.

Ce concept de SEM est étroitement lié à celui de réseau bayésien causal. En effet, on peut montrer que la distribution jointe induite par un SEM est décomposable en produit de distributions conditionnelles associé à un graphe bayésien [5].

Intervention. Intervenir sur une variable X_i consiste à modifier sa distribution conditionnelle p_i en la remplaçant par une nouvelle distribution $\tilde{p}_i$. Cette nouvelle distribution peut toujours dépendre des parents de X_i, auquel cas le lien causal est préservé, ce que l'on qualifie d'intervention douce. Elle peut être également une fonction indépendante des parents de X_i, changeant complètement la distribution ou alors fixant X_i en une valeur. Dans ce cas, la variable devient indépendante de ses parents et l'intervention est qualifiée d'intervention structurelle. Pour illustrer, la distribution jointe $p(\mathbf{X})$ après une intervention structurelle sur la variable X_j et une intervention douce sur X_k s'écrit

$$p(\mathbf{X}) = \prod_{i \neq \{j,k\}} p_i(X_i \,|\, \mathrm{pa}(X_i)) \times \tilde{p}_j(X_j) \times \tilde{p}_k(X_k \,|\, \mathrm{pa}(X_k)).$$

Dans le DAG associé à cette distribution, les arêtes reliant X_k à ses parents seront préservées, tandis que celles reliant X_j à ses parents seront supprimées. En raison de cette modification structurelle, ce nouveau graphe sera désigné sous le nom de graphe mutilé [3, 5].

La notion d'intervention constitue la base de la définition d'un lien causal. En effet, si une variable Y reste dépendante d'une autre variable X malgré les interventions effectuées sur cette dernière, cela signifie qu'elles sont causalement liées. On considère alors que X est la cause de Y et Y la l'effet de X.

Couverture de Markov. La couverture de Markov [6] d'une variable est définie pour un réseau bayésien comme étant le plus petit ensemble qui la rend conditionnellement indépendante de toute autre variable. D'un point de vue causal et sous certaines hypothèses [9, 6], cet ensemble est unique et correspond aux parents, aux enfants et aux conjoints (autres parents des enfants) de cette variable.

3 Définition du problème

Objectif. Nous cherchons à créer un modèle de prédiction qui soit robuste face à de nouvelles observations. C'est à dire un modèle capable de prédire correctement une variable cible avec des données différentes de celles vues à l'entraînement. Avec le formalisme de l'intervention, évoqué plus tôt, il est possible de modéliser ces décalages de distribution. Puisque certaines relations dans la structure causale des données restent inchangées malgré ces variations [7], des nouvelles distributions peuvent néanmoins induire un comportement similaire du modèle.

Un ensemble de variables causalement liées à la variable cible semble pertinent à envisager. La difficulté réside dans la détermination des variables causales assurant une robustesse face aux changements de distribution. La couverture de Markov, théoriquement optimale, est un candidat car elle rend toute autre variable indépendante de la cible. Toutefois, en cas d'intervention affectant l'un de ses enfants, ceux-ci deviennent indépendants, ce qui pourrait nuire à la prédiction. Un autre candidat serait l'ensemble des parents de la cible [2, 7], car ils la déterminent directement. Cependant, cet ensemble pourrait être trop restrictif et réduire la précision lors d'un changement de distribution.

Ce travail vise donc à évaluer empiriquement la capacité prédictive de ces deux candidats. La prochaine partie est consacrée à la présentation de l'algorithme BART (Bayesian Additive Regression Trees, [1]) qui nous servira pour cette comparaison.

Méthode utilisée. BART est une méthode de régression bayésienne non paramétrique basée sur une somme d'arbres de régression binaires. Cette approche permet une inférence bayésienne complète, incluant à la fois des estimations ponctuelles et par intervalles de la fonction de régression inconnue. Pour une meilleure compréhension, nous allons nous attarder sur deux points clés : les arbres de décision binaires ainsi que la distribution a priori utilisée.

Un arbre est paramétré par T, une structure de règles de décision, et par $M = (\mu_1, \ldots, \mu_b)$, des valeurs associées aux b nœuds terminaux (feuilles). En utilisant m pour désigner le nombre d'arbres utilisés, le problème se pose de cette manière

$$Y = f(x) + \epsilon = \sum_{j}^{m} g(x; T_j, M_j) + \epsilon$$

où $\epsilon \sim \mathcal{N}(0, 1)$ et $g(x; T_j, M_j) = \sum_{i=1}^{b} \mu_{i,j} \mathbb{1}_{x \in R_{i,j}}$ avec $R_{i,j}$ le nœud terminal i de l'arbre j.

Une distribution a priori très informative est imposée sur ces derniers, limitant l'information qu'ils peuvent contenir sur la variable d'intérêt. Pendant l'apprentissage, chaque arbre est ajusté sur les résidus laissés par les autres à l'aide de l'algorithme MCMC (Makov Chain Monte Carlo) *bayesian backfitting* [4]. Ainsi, chaque arbre se concentre sur la portion de la variable d'intérêt non expliquée par les autres. L'algorithme ergodique produit des échantillons d'arbres qui convergent vers la distribution a posteriori, permettant d'estimer la fonction désirée.

4 Expériences numériques

Dans cette section, nous analysons la performance prédictive des modèles entraînés sur les variables de la couverture de Markov et sur les parents de la variable cible. Nous les comparons à des modèles entraînés sur deux autres ensembles : un incluant toutes les variables du SEM (Sys) et l'autre avec ces variables plus des variables de bruit indépendantes. Les performances sont comparées selon deux types de prédiction : sur des données test (issues de la même distribution que l'entraînement) et sur des données hors distribution, obtenues après une intervention modifiant la distribution d'origine.

Pour évaluer les performances, nous utilisons le MSE (Mean Square Error) dans le cadre du processus de génération des données décrit dans le tableau 1. Les interventions portent sur les trois types de variables de la couverture de Markov de Y : parents, enfants et conjoints. Nous avons choisi des interventions structurelles en fixant les variables à leur moyenne avant intervention, supposant qu'elles préservent partiellement la cohérence avec la distribution d'origine.

Variable cible			
$Y = 0.5P_1P_2 + 0.2P_3 + \mathcal{N}$			
Parents	$P_1 \sim \mathcal{N}$ $P_2 = 2V_1^3 + \mathcal{N}$ $P_3 = 3V_2 + \mathcal{N}$	**Enfants**	$E_1 = 0.5C_1Y + \mathcal{N}$ $E_2 = 0.3C_2 + Y + \mathcal{N}$ $E_3 = 0.2C_3 + 0.9P_3 + Y^2 + \mathcal{N}$
Conjoints	$C_1 = 0.6V_1 + \mathcal{N}$ $C_2 = 0.2C_1 + \mathcal{N}$ $C_3 = 0.3V_2 + 0.1C_2 + \mathcal{N}$	**Autres**	$V_1, V_2 \sim \mathcal{N}$ $V_3 = 0.3V_2 + \mathcal{N}$ $V_4 = 0.7E_1C_1 + 0.3P_1 + \mathcal{N}$ $V_5 = 0.6E_2 + 0.2V_4 + \mathcal{N}$ $V_6 = 0.3E_3 + 0.4V_5 + \mathcal{N}$

TABLE 1 – Définition des variables du SEM. $\mathcal{N}$ désigne un bruit gaussien centré réduit.

Chaque variable contient $n = 1000$ observations, avec l'ajout de 65 variables indépendantes de bruit. Chaque prédiction est répétée 500 fois. La figure 1 résume les résultats. Les prédictions issues de la distribution test et celles après intervention sur les conjoints sont similaires. Dans ce cas, les meilleurs ensembles de variables incluent toutes celles du SEM et la couverture de Markov de Y, bien que la différence soit faible. Les interventions sur les parents et enfants entraînent des changements plus marqués. Fixer les parents de Y améliore la précision de la couverture de Markov. Intervenir sur les enfants renforce la robustesse des parents, tandis que d'autres variables de la couverture deviennent indépendantes, réduisant la prédiction. Dans l'ensemble, cette intervention rend la prédiction plus difficile pour chaque candidat.

5 Conclusion

Dans cet article, nous avons exploré empiriquement l'impact des interventions sur la robustesse des modèles prédictifs. Nos résultats montrent que la sélection de variables joue un rôle dans la capacité d'un modèle à généraliser sur des données hors distribution. Plus précisément, nous avons mis en évidence que l'ensemble des parents de la variable cible, bien que parfois restrictif, offre une meilleure robustesse face aux interventions sur les enfants, tandis que la couverture de Markov semble mieux préserver la précision prédictive dans des conditions standards. Pour approfondir cette approche, il serait intéressant d'analyser plus en détail l'impact qu'une intervention peut avoir sur une distribution pour mieux comprendre ses effets.

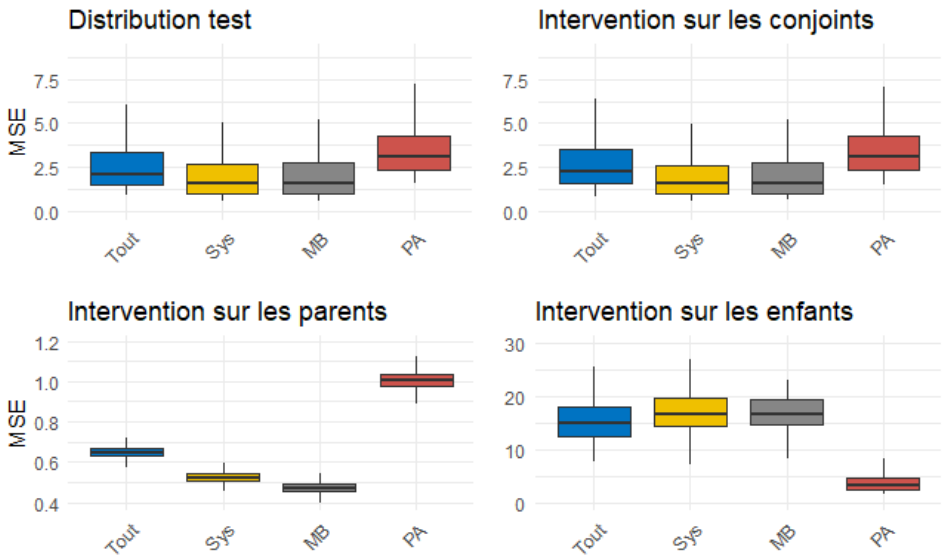


FIGURE 1 – Boxplots des MSE pour les quatre ensembles dans chaque scénario. Nous avons retiré les outliers pour plus de visibilité.

Références

[1] Hugh A CHIPMAN, Edward I GEORGE et Robert E MCCULLOCH. « BART : Bayesian additive regression trees ». In : (2010).

[2] Rune CHRISTIANSEN et al. « A causal framework for distribution generalization ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (2021), p. 6614-6630.

[3] Frederick EBERHARDT et Richard SCHEINES. « Interventions and causal inference ». In : *Philosophy of science* 74.5 (2007), p. 981-995.

[4] Trevor HASTIE et Robert TIBSHIRANI. « Bayesian backfitting (with comments and a rejoinder by the authors ». In : *Statistical Science* 15.3 (2000), p. 196-223.

[5] Judea PEARL et al. « Models, reasoning and inference ». In : *Cambridge, UK : CambridgeUniversityPress* 19.2 (2000), p. 3.

[6] Judea PEARL. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Elsevier, 2014.

[7] Jonas PETERS, Peter BÜHLMANN et Nicolai MEINSHAUSEN. « Causal inference by using invariant prediction : identification and confidence intervals ». In : *Journal of the Royal Statistical Society Series B : Statistical Methodology* 78.5 (2016), p. 947-1012.

[8] Jonas PETERS, Dominik JANZING et Bernhard SCHÖLKOPF. *Elements of causal inference : foundations and learning algorithms*. The MIT Press, 2017.

[9] Peter SPIRTES, Clark GLYMOUR et Richard SCHEINES. *Causation, prediction, and search*. MIT press, 2001.

The Case for Time in Causal DAGs

Alexander G. Reisach [1] & Alberto Suárez [2] & Sebastian Weichwald [3] & Antoine Chambaz [1]

[1] *CNRS, MAP5, Université Paris Cité, France*
[2] *Universidad Autónoma de Madrid, Spain*
[2] *University of Copenhagen, Denmark*

Résumé. Nous proposons d'intégrer explicitement le temps dans la définition des variables des graphes causaux orientés acycliques (DAGs). La causalité exige que les causes précèdent les effets dans le temps, ce qui implique que les relations causales entre variables dans un ordre temporel donné peuvent différer dans un autre. Ainsi, tout modèle causal nécessite une qualification temporelle, même lorsqu'il ne décrit pas une série temporelle de mesures répétées. Nous formalisons une notion de temps pour les variables causales et soutenons qu'elle résout l'ambiguïté existante dans les DAGs causaux et qu'elle est essentielle pour évaluer la validité de l'hypothèse d'acyclicité. Si les variables sont séparées dans le temps, leur relation causale est nécessairement acyclique. Dans le cas contraire, l'acyclicité dépend de l'absence de cycles causaux autorisés par l'ordre temporel. Nous introduisons une distinction formelle entre ces deux conditions et en exposons les implications respectives. Enfin, nous établissons des liens entre notre contribution et divers courants de la littérature sur la causalité et discutons des conséquences de l'intégration du temps pour l'interprétation et l'applicabilité des DAGs en tant que modèles causaux.

Mots-clés. Causalité, DAG, Modèle Graphique, Variable Temporelle, Représentation Causale

Abstract. We make the case for incorporating time explicitly into the definition of variables in causal directed acyclic graphs (DAGs). Causality requires that causes precede effects in time, meaning that the causal relationships between variables in one time order may not be the same in another. Therefore, any causal model requires temporal qualification; this applies even if the model does not describe a time series of repeated measurements. We formalize a notion of time for causal variables and argue that it resolves existing ambiguity in causal DAGs and is essential to assessing the validity of the acyclicity assumption. If variables are separated in time, their causal relationship is necessarily acyclic. Otherwise, acyclicity depends on the absence of causal cycles permitted by the time order. We introduce a formal distinction between these two conditions and lay out their respective implications. We outline connections of our contribution with different strands of the broader causality literature and discuss the ramifications of considering time for the interpretation and applicability of DAGs as causal models.

Keywords. Causality, DAG, Graphical Model, Temporal Variable, Causal Representation

1 Time in Causal DAGs

In contrast to other approaches to causal analysis (e.g. Imbens and Rubin 2015), causal DAGs have been developed explicitly for 'nontemporal statistical data' (Pearl 2009, Chapter 2). The nature of this approach obscures the need for temporal precedence in causation and the rationale for why the acyclicity assumption may or may not be valid. We make the case for a formal and explicit treatment of the time of variables in causal DAGs because time is essential for the interpretation of causal relationships, and because time is crucial to ensuring that the acyclicity assumption underlying the model class is valid. We emphasize that this apply generally, not only to time series of repeated measurements.

A DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a set of vertices $\mathcal{V}$, and a set of directed edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, such that if $(X, Y) \in \mathcal{E}$, then there is a directed edge from X to Y. We call 'directed path' an ordered collection of directed edges such that the end point of the k-th edge is the start point of the $(k+1)$-th edge. By definition, DAGs are acyclic, that is, there are no directed paths that begin and end in the same node. A causal DAG represents an external causal system (Dawid 2010, Section 5). Each vertex corresponds to a random variable, and their joint distribution is assumed to be Markov with respect to the DAG (e.g. Pearl 2009, Definition 1.2.2; Peters et al. 2017, Definition 6.21). A directed edge $(X, Y) \in \mathcal{E}$, also denoted as $X \to Y$, indicates that X is a direct cause of Y within the variables of the graph.

1.1 A Formal Notion of Time for Causal Variables

We introduce a formal treatment of within-sample time for causal variables. For convenience, we will use the following notation: for any $k \in \mathbb{N}^*$, for any set S of cardinality greater than or equal to k, $\mathcal{P}_k(S)$ denotes the collection of subsets of S which are of cardinality k.

Let $\mathcal{T}_x \subset \mathbb{R}_+$ be a countable set of time points, and let $k_x \in \mathbb{N}^*$ be a deterministic integer smaller than or equal to $|\mathcal{T}_x|$. We define a real-valued stochastic process $\mathbb{X} = (\mathbb{X}_t)_{t \in \mathcal{T}_x}$ over these time points, and denote by $\mathbb{T}_x \subseteq \mathcal{T}_x$ a random subset consisting of k_x time points. The support $\mathcal{S}_x$ of the distribution of $\mathbb{T}_x$ is a subset of $\mathcal{P}_{k_x}(\mathcal{T}_x)$. The variables $\mathbb{X}_t$ are measured for every $t \in \mathbb{T}_x$.

Let $f_x \colon \mathbb{R}^{k_x} \to \mathbb{R}$ be a real-valued deterministic aggregation function, so that $X = f_x((\mathbb{X}_t)_{t \in \mathbb{T}_x})$ is an aggregate variable of the measurements $(\mathbb{X}_t)_{t \in \mathbb{T}_x}$, which form its temporal components. We impose further that if $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{k_x}$ are such that there exists $t \in \{1, \ldots, k_x\}$ for which $\mathbf{x}_s = \mathbf{x}'_s$ for every $s \neq t$ but $\mathbf{x}_t \neq \mathbf{x}'_t$, then $f_x(\mathbf{x}) \neq f_x(\mathbf{x}')$. This is to ensure that an intervention on one temporal component also results in a change in the corresponding aggregate variable. For instance, f_x cannot act as a component selector and it cannot multiply variables if one of them is equal to zero. To emphasize how the aggregate variable X inherits its temporal structure from the underlying stochastic process, we introduce the notation

$$X_{\mathbb{T}_x} \coloneqq f_x\left((\mathbb{X}_t)_{t \in \mathbb{T}_x}\right), \tag{1}$$

which captures that a causal variable can be seen as the aggregate of one or multiple temporal components. Owing to the essential role of time in causality, we choose to keep the time

index explicit even after the aggregation into a single real-valued random variable. Moving forward, all causal variables will be of the form given in Equation (1), and we assume the existence of an underlying 'atomic' causal DAG between all time point-specific components of the stochastic processes involved in the phenomenon of interest. From the atomic causal DAG it is possible to derive a directed graph between these aggregate variables, which may or may not be acyclic. For a detailed treatment of inference rules over macro-variables, see Anand et al. (2023). Denoting the support of the joint distribution of two sets of time points $(\mathbb{T}_x, \mathbb{T}_y)$ as

$$\mathcal{S}_{xy} \subseteq \mathcal{S}_x \times \mathcal{S}_y \subseteq \mathcal{P}_{k_x}(\mathcal{T}_x) \times \mathcal{P}_{k_y}(\mathcal{T}_y), \tag{2}$$

we say that $X_{\mathbb{T}_x}$ causes another causal variable $Y_{\mathbb{T}_y}$ if there exist $(t_x, t_y) \in \mathcal{S}_{xy}$ such that $\mathbb{X}_{t_x}$ causes $\mathbb{Y}_{t_y}$ in the atomic DAG (we do not consider self-loops on the level of aggregate variables, e.g. if $\mathbb{X}_{t_{x_1}}$ causes $\mathbb{X}_{t_{x_2}}$ for some $t_{x_1}, t_{x_2} \in \mathcal{S}_x$). Our formalization of the time associated with causal variables enables us to determine the possibility of causal cycles and to state precise requirements for acyclicity in terms of the atomic causal DAG.

Example 1 (separation in time). Consider a medical treatment-health setting following our formalism. Let $X_{\mathbb{T}_x}$ be an aggregate of $k_x = 2$ treatment measurements in $(\mathbb{X}_t)_{t\in\mathcal{T}_x}$, where $\mathcal{T}_x = \{0, 0.4\}$. Let $Y_{\mathbb{T}_y}$ be an aggregate variable of $k_y = 2$ health measurements in $(\mathbb{Y}_t)_{t\in\mathcal{T}_y}$, where $\mathcal{T}_y = \{0.6, 1\}$. It follows that $\mathbb{T}_x = \mathcal{T}_x$ and $\mathbb{T}_y = \mathcal{T}_y$. Thus, the only possible combination of time points is $\mathcal{S}_{xy} = \mathcal{P}_2(\mathcal{T}_x) \times \mathcal{P}_2(\mathcal{T}_y) = \{(\{0, 0.4\}, \{0.6, 1\})\}$. Let the aggregation functions f_x and f_y take the sum of their inputs, that is, $X_{\mathbb{T}_x} := \sum_{t\in\mathbb{T}_x} \mathbb{X}_t$ and $Y_{\mathbb{T}_y} := \sum_{t\in\mathbb{T}_y} \mathbb{Y}_t$. Such a setting may correspond to patients always receiving a treatment at both of two possible time points, followed always by a health screening at both of two possible time points.

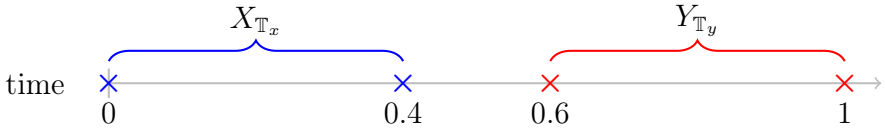
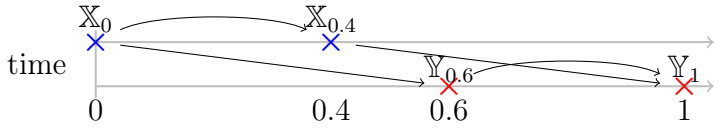
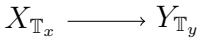


Figure 1: Two aggregate causal variables that are separated in time.

Figure 1 shows the timeline, time points of the stochastic process $(\mathbb{X}_t)_{t\in\mathcal{T}_x}$ (blue ×) and $(\mathbb{Y}_t)_{t\in\mathcal{T}_y}$ (red ×), and their aggregation into the aggregate variables.



(a) Atomic causal DAG.



(b) Causal DAG between the aggregate variables.

Figure 2: An atomic causal DAG and aggregate causal DAG compatible with Example 1.

Figure 2 shows an atomic causal DAG (Figure 2a) that is compatible with the example, and its corresponding causal DAG between the aggregate variables (Figure 2b). The temporal

components of each aggregate variable are drawn on separate (gray) time arrows belonging to the same within-sample timeline, as is indicated by the vertical connection. The solid black arrows show causal relationships.

2 A Tale of Two Acyclicities

Acyclicity of the causal relationship between two variables $X_{\mathbb{T}_x}$ and $Y_{\mathbb{T}_y}$ can be defined in terms of the causal relationships between the temporal components of the variables, that is, the atomic DAG.

Definition 1 (Acyclicity)**.** *Acyclicity between two causal variables $X_{\mathbb{T}_x}$ and $Y_{\mathbb{T}_y}$ holds if and only if in the atomic causal DAG*

$$\begin{aligned} &\exists (s_x, s_y) \in \mathcal{S}_x \times \mathcal{S}_y, \exists t_x \in s_x, \exists t_y \in s_y, \quad (\mathbb{X}_{t_x} \rightsquigarrow \mathbb{Y}_{t_y}) \\ \implies &\forall (s'_x, s'_y) \in \mathcal{S}_x \times \mathcal{S}_y, \forall t'_x \in s'_x, \forall t'_y \in s'_y, \neg(\mathbb{Y}_{t'_y} \rightsquigarrow \mathbb{X}_{t'_x}). \end{aligned}$$

We further call a graph $\mathcal{G}$ acyclic if acyclicity holds for all pairs of variables in $\mathcal{G}$.

2.1 Acyclicity I: By Time Order

A sufficient condition for an acyclic causal relationship between two variables $X_{\mathbb{T}_x}$ and $Y_{\mathbb{T}_y}$ is satisfied if all time points of one variable always precede those of the other variable. We denote the temporal precedence of $X_{\mathbb{T}_x}$ with respect to $Y_{\mathbb{T}_y}$ by the shorthand

$$\mathbb{T}_x \prec \mathbb{T}_y \coloneqq \forall (s_x, s_y) \in \mathcal{S}_{xy}, \max(s_x) < \min(s_y) \tag{3}$$

(the order relation $\prec$ is transitive), and define 'time-acyclicity' as follows.

Definition 2 (Time-acyclicity)**.** *Time-acyclicity of two causal variables $X_{\mathbb{T}_x}$ and $Y_{\mathbb{T}_y}$ holds if and only if $(\mathbb{T}_x \prec \mathbb{T}_y)$ and/or $(\mathbb{T}_y \prec \mathbb{T}_x)$. We further call a graph $\mathcal{G}$ time-acyclic if time-acyclicity holds for all pairs of variables in $\mathcal{G}$.*

Example 1 exhibits time-acyclicity by Definition 2. Time-acyclicity can be complex if the the measurement times of some variables depend on those of others.

2.2 Acyclicity II: By Absence of Causal Cycles

Discussions of the acyclicity assumption in the causal DAG literature typically focus on the presence or absence of causal cycles such as feedback loops (see e.g. Brouillard et al. 2024). However, given that causes precede effects, causal cycles can only occur if there is no time-acyclicity (Definition 2). We term acyclicity due to the absence of causal cycles where they are possible 'effect-acyclicity' and define it as follows.

Definition 3 (Effect-acyclicity)**.** *Effect-acyclicity holds between two causal variables $X_{\mathbb{T}_x}$ and $Y_{\mathbb{T}_y}$ if and only if there is acyclicity by Definition 1 but no time-acyclicity by Definition 2. We further call a graph $\mathcal{G}$ effect-acyclic if effect-acyclicity holds for all pairs of variables in $\mathcal{G}$.*

Example 2 (effect-acyclicity). Consider another variation of our running medical treatment-health example. Let $X_{\mathbb{T}_x}$ be an aggregate of $k_x = 2$ treatment measurements in $(\mathbb{X}_t)_{t\in\mathcal{T}_x}$ where $\mathcal{T}_x = \{0, 0.6\}$. Let $Y_{\mathbb{T}_y}$ be an aggregate of $k_y = 2$ health measurements in $(\mathbb{Y}_t)_{t\in\mathcal{T}_y}$, where $\mathcal{T}_y = \{0.4, 1\}$. It follows that $\mathbb{T}_x = \mathcal{T}_x$, $\mathbb{T}_y = \mathcal{T}_y$, and $\mathcal{S}_{xy} = \mathcal{P}_2(\mathcal{T}_x) \times \mathcal{P}_2(\mathcal{T}_y) = \{(\{0, 0.6\}, \{0.4, 1\})\}$. Let the aggregation functions f_x and f_y each take the sum of their inputs. Consider the case of patients who receives treatment at time $t = 0$, which affects health at time 0.4, which affects treatment at time 0.6, which in turn affects health at time 1.

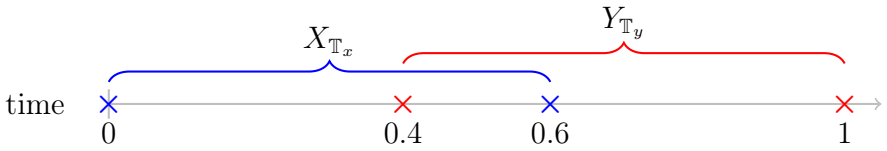


Figure 3: An illustration of Example 2, where the time order permits causal cycles.

The overlap of the red and blue brackets indicates a lack of time order by Equation (3) between $X_{\mathbb{T}_x}$ and $Y_{\mathbb{T}_y}$, as can be read off $\mathcal{S}_{xy}$. The acyclicity assumption can only hold by effect-acyclicity (Definition 3), otherwise the variables cannot be validly represented in a causal DAG.

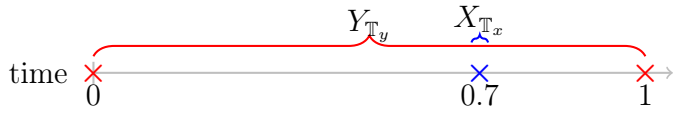
3 The Acyclicity Assumption

The acyclicity assumption is a defining characteristic of DAGs. For a DAG to be a fitting model of the causal relationships between some variables, the acyclicity assumption must be valid for these variables. The concepts of time-acyclicity and effect-acyclicity can be used for assessing this validity. Assessing time-acyclicity is principally an analysis of the measurement procedure, whereas assessing effect-acyclicity requires additionally examining potential causal cycles. The explicit role of time in our formalism clarifies the practical feasibility of such an assessment by specifying measurement times, thereby highlighting acyclicity as a testable assumption.

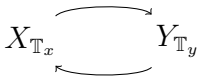
It is often suggested that even cyclic causal relationships may be acyclic at a sufficiently fine-grained temporal resolution (e.g. Pearl 2009, Chapter 1.4.1; Peters et al. 2017, Chapter 2.3.3). Using our formalism, we give a simple example of how a cyclic effect can be unrolled in time to obtain a DAG.

Example 3 (unrolling in time). Consider another variation of our running medical treatment-health example. Let $X_{\mathbb{T}_x}$ be an aggregate of $k_x = 1$ health measurement in $(\mathbb{X})_{t\in\mathcal{T}_x}$, where $\mathcal{T}_x = \{0.7\}$. Let $Y_{\mathbb{T}_y}$ be an aggregate of $k_y = 2$ treatment measurements in $(\mathbb{Y}_t)_{t\in\mathcal{T}_y}$, where $\mathcal{T}_y = \{0, 1\}$. It follows that $\mathbb{T}_x = \mathcal{T}_x$, $\mathbb{T}_y = \mathcal{T}_y$, and $\mathcal{S}_{xy} = \{(\{0.7\}, \{0, 1\})\}$, meaning that measurements are always taken at all time points. Let the aggregation function f_x be the identity, and let f_y take the sum of its inputs. Assume that this setting corresponds to an initial health measurement that informs a treatment, which leads to a final health

measurement. Assume further that initial health also affects final health directly. This is to say that $\mathbb{Y}_0$ causes $\mathbb{X}_{0.7}$, $\mathbb{X}_{0.7}$ causes $\mathbb{Y}_1$, and $\mathbb{Y}_0$ causes $\mathbb{Y}_1$ in the atomic DAG. The temporal



(a) Timeline and aggregation; no time-acyclicity by Definition 2.

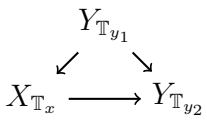


(b) Cyclic causal directed graph.

Figure 4: The timeline, atomic causal DAG, and causal directed graph of Example 3, where the variables $X_{\mathbb{T}_x}$ and $Y_{\mathbb{T}_y}$ have a cyclic causal relationship.

aggregation (Figure 4a) gives rise to a causal cycle (Figure 4b). The causal cycle can be resolved if the health variable is split into two parts: one representing health before the treatment, and one after. The result can be seen in Figure 5. Splitting $Y_{\mathbb{T}_y}$ into $Y_{\mathbb{T}_{y_1}}$ and $Y_{\mathbb{T}_{y_2}}$,

(a) New causal variables with time-acyclicity by Definition 2.



(b) New causal DAG.

Figure 5: Resolving the causal cycle by splitting $Y_{\mathbb{T}_y}$ in time.

as shown in Figure 5a, transforms the cyclic causal relationship into the confounder-DAG between the new variables shown in Figure 5b.

4 Summary: Toward More Applicable Causal DAGs.

Taking into account the role of time means tackling the question of the applicability of causal DAGs to real-world data in earnest. Before the question "what is the right DAG?", comes the question "is there a DAG?". The crucial step in applying causal DAGs consists in defining meaningful variables that allow for causal DAGs, rather than treating a set of variables as given and reasoning about *the* DAG between them. An explicit notion of time facilitates finding such variables and ensuring the acyclicity assumption is met, for example by establishing measurement protocols that satisfy time-acyclicity. Our formalism allows for incorporating time in a principled fashion and opens many questions for future research. For example, it will be interesting to examine the change of causal effects with time distance and to what extent some time points may be representative of others. Moreover, further research is required to analyze the implications in terms of faithfulness violations, and to apply our notion of time to other classes of causal graphical models. Although our contribution shows the scope of causal DAGs to be more specific than it may otherwise appear, it also facilitates their discovery, clarifies their interpretation, and leaves them a less metaphysical and all the more practical tool for causal analysis.

References

[1] Tara V. Anand, Adele H. Ribeiro, Jin Tian, and Elias Bareinboim. Causal Effect Identification in Cluster DAGs. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 10. 2023, pp. 12172–12179 (cit. on p. 3).

[2] Philippe Brouillard, Chandler Squires, Jonas Wahl, Konrad P. Kording, Karen Sachs, Alexandre Drouin, and Dhanya Sridhar. The Landscape of Causal Discovery Data: Grounding Causal Discovery in Real-World Applications. In: *arXiv preprint* (2024) (cit. on p. 4).

[3] Philip A. Dawid. Beware of the DAG! In: *Causality: Objectives and Assessment*. PMLR. 2010, pp. 59–86 (cit. on p. 2).

[4] Guido W. Imbens and Donald B. Rubin. Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press, 2015 (cit. on p. 2).

[5] Judea Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2009 (cit. on pp. 2, 5).

[6] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Elements of Causal Inference: Foundations and Learning Algorithms. MIT Press, 2017 (cit. on pp. 2, 5).

Inférence causale 2

A Generalized Propensity Score Estimation Methodology for Discrete and Continuous Treatments

Felipe L. Angelim Vieira[1] & Alessandro Leite[2]

[1]*Mercado Libre, Brazil, felipeangelim@gmail.com*
[2]*INSA Rouen Normandie, Normandie Univ, LITIS UR 4108, 76000 Rouen, France*

Résumé. Les scores de propension (SP) constituent une stratégie efficace pour contrôler les biais lors de l'estimation des effets causaux à partir de données observationnelles. Néanmoins, les méthodes traditionnelles rencontrent souvent des difficultés avec des structures de traitements complexes, allant au-delà des traitements binaires ou discrets. Dans ce contexte, ce travail propose une méthodologie d'estimation des scores de propension qui prend en compte des structures de traitements généralisées, sans imposer d'hypothèses paramétriques restrictives. Nous adoptons un modèle de classification probabiliste pour estimer les poids stabilisés et les scores de propension via une estimation du rapport de densité, rendant notre méthode adaptable à diverses formes de traitements. Les résultats expérimentaux montrent des améliorations notables en termes de précision et de stabilité sur différents ensembles de données synthétiques et semi-synthétiques, même lorsque le nombre de covariables augmente.

Mots-clés. Score de propension, Traitements continus, Estimation dose-réponse, Effet moyen du traitement.

Abstract. Propensity scores (PS) comprise an effective strategy to control for biases when estimating causal effects based on observational data. Nevertheless, traditional methods often struggle with complex treatment structures beyond binary or discrete treatments. In this context, this work introduces a methodology for propensity score estimation that accommodates generalized treatment structures without imposing restrictive parametric assumptions. We follow a probabilistic classification model to estimate the stabilized weights and the propensity scores via density ratio estimation, making our method adaptable to various treatment forms. Experimental results show notable performance improvements in accuracy and stability across different synthetic and semi-synthetic datasets, even as the number of covariates increases. This suggests that our approach is flexible enough to be considered for challenging real-world applications when estimating causal effects, such as healthcare, policy analysis, and personalized marketing.

Keywords. Propensity Score, Continuous treatments, dose-response estimation, average treatment effect

1 Introduction

Causal inference is fundamental in both industrial and research applications that rely on observational data to answer counterfactual questions [1, 2]. In the absence of randomized control trials, observational data can provide valuable causal insights, provided that appropriate causal methods are applied [3]. However, treatment assignment in observational data is often influenced by covariates, and failing to account for this bias may lead to misleading conclusions. A widely used approach to mitigate this issue is the propensity score (PS), which represents the probability of assigning a treatment value given the covariates. Propensity scores can be employed to re-weight dataset samples or replace covariates as regressors when estimating response functions. While probabilistic classification models are effective for estimating propensity scores in discrete treatment settings, continuous and mixed treatment regimes typically require additional parametric assumptions.

This work introduces a novel methodology to approximate the PS independent of the treatment structure. The proposed approach can be applied with any binary classification model or integrated as a loss function in neural network architectures, such as VCNet [4]. Through experimental analysis, we demonstrate the efficacy of our method in handling both discrete and continuous treatments, achieving promising results. Furthermore, the generalization capacity of the methodology enhances its applicability for practitioners aiming to build scalable solutions for obtaining causal insights.

2 Problem statement

We consider a dataset $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^n$, consisting of i.i.d. samples from an observational study where treatment assignment depends on covariates. Each sample $(X, T, Y) \in \mathcal{X} \times \mathcal{T} \times \mathcal{Y}$ represents covariates, treatment level, and observed outcome, respectively, drawn from the joint distribution $P_s(Y \mid X, T)P_s(T \mid X)P_s(X)$. Using Rubin's potential outcome framework [5], we define $Y^\star(t)$ as the hypothetical outcome if treatment t were applied, irrespective of the actual treatment received. However, only $y_i = Y^\star(t_i)$ is observed for each sample, making it impossible to directly estimate $Y^\star(t')$ for $t' \neq t_i$. Additionally, observational studies face bias due to treatment assignments influenced by x_i, necessitating methods to address these biases. The goal is to estimate the average dose-response function (ADRF), $\mu(t)$, which captures how the average outcome varies with treatment levels $t \in \mathcal{T}$. It is given by:

$$\mu(t) = \mathbb{E}_{Y' \sim P_s(Y^\star(t))}[Y'] = \mathbb{E}_{X \sim P_s(X)}\mathbb{E}_{Y' \sim P_s(Y^\star(t)|X)}[Y'].$$

under the positivity and ignorability standard causal inference assumptions.

3 Algorithm for estimating propensity score

The core idea of the proposed methodology is to build a synthetic binary classification dataset and to obtain estimates for the density ratio, which can be equal to the generalized propensity score (GPS) or stabilized weights depending on the construction of the dataset. It is deeply

connected to covariate shift methods in machine learning that use classifiers to approximate importance sampling weights in which a model is trained to discriminate between the source and the target distributions P_s and P_t and use the estimated probabilities to approximate the density ratio.

Lemma 1 *Let $n_t, n_s \in \mathbb{N}$. Consider the joint distribution of features $Z \in \mathcal{Z}$ and labels $Q(Z, C) = Q(Z|C)Q(C)$, where $C \sim Bern(\frac{n_t}{n_t+n_s})$. In addition, let class 0 represent the source distribution $Q(Z|C = 0) = P_s(Z)$ and class 1 represent the target $Q(Z|C = 1) = P_t(Z)$. The density ratio $\rho(Z) = \frac{P_t(Z)}{P_s(Z)}$ equals the ratio of conditional class probabilities, adjusted by the prior ratio:*

$$\rho(Z) = \frac{Q(C = 1 \mid Z)}{Q(C = 0 \mid Z)} \frac{n_s}{n_t} \tag{1}$$

In the case of causal inference, the same methodology can be applied by considering the distribution of the observational data set as a source and as a target, a distribution where the assignment of treatment is independent of the covariates. Letting $\mathcal{Z} = \mathcal{X} \times \mathcal{T}$, the following corollary of Lemma 1 shows how forcing P_t to have independence of treatment and covariates leads to a useful ratio of probabilities of receiving a given treatment.

Corollary 2 *Let $X \in \mathcal{X}$ be covariates, $T \in \mathcal{T}$ a continuous treatment, $P_s(Z) = P_s(X, T) = P_s(T|X)P_s(X)$ the data generating process of an observational experiment and $P_t(Z) = P_s(X)P_{cf}(T)$. Then the density ratio $\rho(X, T)$ is:*

$$\rho(X, T) = \frac{P_{cf}(T)}{P_s(T|X)}$$

Furthermore, creating a binary classification dataset according to Lemma 1 leads to:

$$\rho(X, T) = \frac{P_{cf}(T)}{P_s(T|X)} = \frac{Q(C = 1|X, T)}{Q(C = 0|X, T)} \frac{n_s}{n_t} \tag{2}$$

In practice, the treatment probability $P_{cf}(T)$ can be set to any probability we can sample from. The Theorem 3 states how specific choices of P_{cf} lead us to the GPS and stabilized weights.

Theorem 3 *If $P_{cf}(T) = \mathcal{U}(\mathcal{T})$ and we account for the volume of treatment space, then the density ratio estimates the inverse of the propensity score:*

$$\rho(X, T) = \frac{Q(C = 1|X, T)}{Q(C = 0|X, T)} \frac{n_s}{n_t} = \frac{\mathcal{U}(\mathcal{T})}{P_s(T|X)} = \frac{1}{|\mathcal{T}|} \frac{1}{P_s(T|X)} = \frac{w(X, T)}{|\mathcal{T}|} \tag{3}$$

In addition, if we let $P_{cf}(T) = P_s(T)$, we obtain the stabilized weights:

$$\rho(X, T) = \frac{Q(C = 1|X, T)}{Q(C = 0|X, T)} \frac{n_s}{n_t} = \frac{P_s(T)}{P_s(T|X)} = w_s(X, T) \tag{4}$$

While the source distribution corresponds to our observational dataset $\mathcal{D}_s = \{(x_i, t_i)\}$, the target distribution can be constructed synthetically $\mathcal{D}_t = \{(x_i, t'_i) \ : \ x_i \sim P_s(X), \ t'_i \sim P_{cf}(T)\}_{i=1}^{n_r}$. When estimating the propensity score, we assign treatment values sampled from a uniform distribution to samples x_i and, to estimate the stabilized weight, we randomly sample the treatment from the treatment values in $\mathcal{D}_s$. A classifier can then be trained to discriminate whether a sample comes from $\mathcal{D}_s$ or $\mathcal{D}_t$.

3.1 Losses for propensity score estimation

One concern when estimating classification probability to obtain density ratio estimation may be the convergence guarantees we have. Fortunately, as demonstrated by Menon and Ong [6], minimizing any Class Probability Estimation (CPE) loss is equivalent to minimizing a Bregman divergence to the true density ratio, with different composite proper losses focusing on different ranges of the density ratio. In other words, the choice of CPE loss dictates a weight function that describes the cost of density ratio error. The least-squares importance fitting (LSIF) loss has the desired property of uniformly weighting the error of density ratio estimation. This loss, when applied to the proposed classification problem in Theorem 3, is defined as:

$$LSIF(\hat{\rho}) = \mathbb{E}_{P_s(X,T)}(\rho(X,T) - \hat{\rho}(X,T))^2 \tag{5}$$
$$= \mathbb{E}_{P_s(X,T)}\hat{\rho}^2(X,T) - 2\mathbb{E}_{P_s(X)P_{cf}(T)}\hat{\rho}(X,T) + C_\rho \tag{6}$$
$$= \ell_{lsif}(\hat{\rho}) + C_\rho \tag{7}$$

where ρ is the true density ratio and C_ρ is a constant that can be ignored when optimizing for $\hat{\rho}$, since:

$$\rho = \arg\min_{\hat{\rho}} LSIF(\hat{\rho}) = \arg\min_{\hat{\rho}} \ell_{lsif}(\hat{\rho}) \tag{8}$$

The term $\ell_{lsif}(\hat{\rho})$ can be approximated using the observational dataset $\mathcal{D}_s$ by:

$$\ell_{lsif}(\hat{\rho}) \approx \frac{1}{|\mathcal{D}_s|}\sum_{(x_i,t_i)\in\mathcal{D}_s}\hat{\rho}^2(x_i,t_i) - 2\frac{1}{|\mathcal{D}_s|}\sum_{\substack{x_i\in\mathcal{D}_s \\ t_i'\sim P_{cf}(T)}}\hat{\rho}(x_i,t_i') \tag{9}$$

This loss can be easily incorporated into the training loop of neural network models that estimate balancing weights, such as VCNet [4]. Similarly, the logistic loss can be used to approximate the weights:

$$\ell_{logistic}(\hat{\rho}) \approx -\frac{1}{|\mathcal{D}_s|}\sum_{(x_i,t_i)\in\mathcal{D}_s}\log\hat{p}(x_i,t_i) - \frac{1}{|\mathcal{D}_s|}\sum_{\substack{x_i\in\mathcal{D}_s \\ t_i'\sim P_{cf}(T)}}\log\left(1-\hat{p}(x_i,t_i)\right) \tag{10}$$

where, $\hat{p}(x_i,t_i) := (\hat{\rho}(x_i,t_i)+1)^{-1}$. When $P_{cf}(T)$ is set to $P_s(T)$ **we call *stabilized* LSIF (LSIF-S)** or **logistic loss**, and ***uniform*** when $P_{cf}(T)$ is set to $\mathcal{U}(\mathcal{T})$.

4 Experimental Results

We evaluate the effectiveness of the proposed propensity score estimation approach in both continuous and binary treatment settings. To quantify performance, we use the expected mean squared error (EMSE) to compare the estimated function $\hat{\mu}(T)$ with the true function $\mu(T)$. This metric is computed over either the observational treatment distribution $P_s(T)$ or a uniform distribution $\mathcal{U}(\mathcal{T})$, offering insights into model performance across different treatment distributions. For continuous treatments, we employ the GPS method [7] and the VCNet architecture with targeted regularization [4] to estimate the average dose-response function. For binary treatments, we use GPS, standard Inverse Propensity Score Weighting (IPW), and Doubly Robust (DR) estimation. Each method is evaluated using both the proposed approach and baseline propensity score methodologies. In the continuous treatment setting,

we adopt the piecewise linear interpolation method from [4] as a baseline for propensity score estimation, alongside Gradient Boosting and a feed-forward neural network classifier for binary treatment datasets. Our implementation employs feed-forward neural networks trained with the losses defined in Equations (9) and (10), in both uniform and stabilized forms. We pre-process treatment variables for continuous treatments using a spline basis of degree 2 with 10 knots uniformly placed between 0.1 and 0.9 instead of using the raw treatment values.[1] To evaluate the versatility of our methodology, we also use tree-based models, such as the Gradient Boosting Classifier, on synthetic datasets generated according to Equations (3) and (4). A challenge with tree models is their tendency, depending on hyperparameters, to ignore the treatment variable, making predictions solely based on covariates and returning constant class probabilities. If this occurs, we resample treatments from $P_{cf}(T)$ and refit the classifier, repeating up to five iterations until the model produces non-constant probabilities.

Continuous Treatment: Table 1 summarizes the results for the continuous treatment settings. Focusing first on neural network PS estimators, our GPS methodology consistently outperforms the baseline in all cases except for LRS in the Synthetic B dataset. For Synthetic A, all neural network-based PS methods achieved at least a 37% reduction in the EMSE compared to the GPS baseline and a 33% reduction relative to the uniform distribution (EMSE Uniform). Regarding the VCNet + TR estimator, the impact of density ratio estimation varies across datasets. For example, in the Synthetic A dataset with the LSIF loss, it worsens performance, while in other cases, it improves results. Notably, LSIFS consistently surpassed the baseline across all scenarios, demonstrating its robustness. For tree-based classifiers, while the standard Gradient Boosting Classifier (GBC) showed suboptimal performance, its stabilized variant improved errors over the baseline in all datasets, highlighting its effectiveness in this context.

Discrete Treatment: Table 2 describes the results for the proposed propensity score approach in the discrete treatment setting. Except for LRS and GBC, we observe performance improvements over both the neural network and tree-based baselines. To investigate the high error observed with IPW-BR, we found that the neural network struggles to improve accuracy on the validation set, quickly reaching the patience threshold and returning an underfitted model. In contrast, for the other methods, the loss improves for a longer duration before plateauing, though they eventually reach a similar validation loss level.

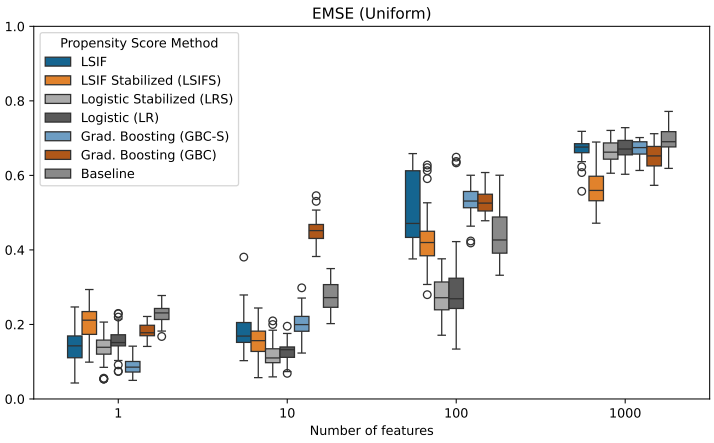


Figure 1: Results for Synthetic A data with varying feature dimensions and a fixed sample size of 1000 using the GPS estimation procedure.

Sensitivity to feature dimension: Figure 1 highlights that although the EMSE increases

[1]Spline pre-processing allows for greater flexibility in capturing non-linear relationships.

Table 1: Comparing methods' performance in the continuous settings

PS Method	GPS		VCNet TR	
	$\sqrt{\text{EMSE}}$	$\sqrt{\text{EMSE (Uniform)}}$	$\sqrt{\text{EMSE}}$	$\sqrt{\text{EMSE (Uniform)}}$
Synthetic A (1000 samples, 10 features)				
BR	0.2845 ± 0.0469	0.2725 ± 0.0373	0.1189 ± 0.0519	**0.1148 ± 0.0556**
LR	0.1293 ± 0.0277	0.1282 ± 0.0261	0.1266 ± 0.0751	**0.1241 ± 0.0674**
LRS	0.1208 ± 0.0355	**0.1181 ± 0.0319**	0.1458 ± 0.0652	0.1458 ± 0.0604
LSIF	**0.1768 ± 0.0543**	0.1807 ± 0.0542	0.4315 ± 0.1628	0.4353 ± 0.1626
LSIF-S	0.1495 ± 0.0394	0.1563 ± 0.0419	0.1061 ± 0.0387	**0.1055 ± 0.0379**
GBC	0.4384 ± 0.0410	0.4520 ± 0.0340	-	-
GBC-S	0.2016 ± 0.0381	0.2017 ± 0.0353	-	-
Synthetic B [8]				
BR	0.0955 ± 0.0137	0.1220 ± 0.0196	0.1144 ± 0.0405	**0.0781 ± 0.0164**
LR	**0.0577 ± 0.0146**	0.0711 ± 0.0214	0.0740 ± 0.0184	0.0698 ± 0.0154
LRS	0.1571 ± 0.0445	0.2423 ± 0.0719	**0.0707 ± 0.0170**	0.0811 ± 0.0265
LSIF	**0.0598 ± 0.0155**	0.0772 ± 0.0235	0.0736 ± 0.0185	0.0784 ± 0.0206
LSIF-S	0.0891 ± 0.0267	0.1181 ± 0.0491	0.1077 ± 0.0286	**0.0724 ± 0.0183**
GBC	**0.1649 ± 0.0451**	0.2410 ± 0.0697	-	-
GBC-S	**0.0707 ± 0.0195**	0.0912 ± 0.0366	-	-
News (semi-synthetic)				
BR	0.1762 ± 0.0741	0.1556 ± 0.0545	0.1539 ± 0.0614	**0.1346 ± 0.0426**
LR	0.1320 ± 0.0672	**0.1045 ± 0.0507**	0.1410 ± 0.0583	0.1285 ± 0.0494
LRS	0.1299 ± 0.0703	**0.1026 ± 0.0511**	0.1421 ± 0.0531	0.1314 ± 0.0407
LSIF	0.1220 ± 0.0592	**0.1022 ± 0.0432**	0.1448 ± 0.0601	0.1392 ± 0.0534
LSIF-S	0.1293 ± 0.0714	**0.1016 ± 0.0513**	0.1334 ± 0.0653	0.1249 ± 0.0649
GBC	0.1606 ± 0.0538	**0.1487 ± 0.0450**	-	-
GBC-S	0.1510 ± 0.0689	**0.1182 ± 0.0530**	-	-

BR: baseline, **LR**: logistic regression, **LRS**: logistic regression stabilized, **LSIF** (our): least-squares importance fitting
LSIF-S (our): LSIF stabilized, **GBC-S**: Gradient Boosting Classifier (stabilized), **GBC** (our): Gradient Boosting Classifier

with the dimensionality, all neural network methods consistently outperform the baseline, except for the LSIF loss at 100 features. Logistic losses, in particular, show significantly lower errors compared to the baseline across all feature dimensions. For gradient-boosting trees, the results indicate that while GBC-S performs well in low-dimensional settings, its performance deteriorates as the number of features increases. Additionally, the performance of GBC suggests that it is not well-suited for density ratio estimation in high-dimensional problems.

5 Conclusion

This study introduces a novel methodology for estimating propensity scores that supports diverse treatment structures, including continuous and binary regimes. Experimental results demonstrate that the proposed approach consistently outperforms state-of-the-art methods, even in high-dimensional covariate settings. These findings have significant implications for both scientific and industrial applications, such as advancing causal analysis in healthcare studies of drug dosages or improving causal insights from complex observational data in industry. Despite these promising results, several challenges remain. For example, tree-based classifiers may require careful pre-processing of the joint covariate-treatment space to

Table 2: Comparing the different methods' performance on the News 1D-Discrete dataset

PS Method	GPS		IPW		DR	
	$\sqrt{\text{EMSE}}$	$\sqrt{\text{EMSE}}$ (Uniform)	$\sqrt{\text{EMSE}}$	$\sqrt{\text{EMSE}}$ (Uniform)	$\sqrt{\text{EMSE}}$	$\sqrt{\text{EMSE}}$ (Uniform)
BR	0.0975 ± 0.0426	0.1330 ± 0.0613	0.5087 ± 0.0260	0.7156 ± 0.0081	0.0671 ± 0.0545	**0.0441 ± 0.0242**
BTC	0.1366 ± 0.0632	0.1824 ± 0.0894	0.3955 ± 0.0528	0.5559 ± 0.0732	**0.0284 ± 0.0110**	0.0356 ± 0.0182
LR	0.0833 ± 0.0454	0.1125 ± 0.0646	**0.0241 ± 0.0098**	0.0260 ± 0.0171	0.0348 ± 0.0129	0.0388 ± 0.0218
LRS	0.1215 ± 0.0366	0.1691 ± 0.0515	0.0459 ± 0.0181	0.0612 ± 0.0269	**0.0373 ± 0.0172**	0.0388 ± 0.0229
LSIF	0.0865 ± 0.0451	0.1146 ± 0.0659	**0.0269 ± 0.0123**	0.0308 ± 0.0207	0.0349 ± 0.0135	0.0393 ± 0.0227
LSIF-S	0.0870 ± 0.0393	0.1199 ± 0.0551	**0.0279 ± 0.0057**	0.0333 ± 0.0101	0.0340 ± 0.0134	0.0379 ± 0.0225
GBC	0.0663 ± 0.0493	0.0876 ± 0.0710	0.1007 ± 0.0609	0.1371 ± 0.0851	0.0551 ± 0.0352	**0.0458 ± 0.0283**
GBC-S	0.0595 ± 0.0410	0.0758 ± 0.0611	0.0434 ± 0.0183	0.0575 ± 0.0266	0.0410 ± 0.0233	**0.0391 ± 0.0236**

BR: baseline, **LR**: logistic regression, **LRS**: logistic regression stabilized, **LSIF** (our): least-squares importance fitting, **LSIF-S** (our): LSIF stabilized, **BTC**: Baseline Tree classifier, **GBC-S**: Gradient Boosting Classifier (stabilized), **GBC** (our): Gradient Boosting Classifier

balance essential interaction modeling and overfitting prevention. Furthermore, the current methodology does not extend to structured data inputs (e.g., images or text) or sequential and multi-dimensional treatments. Future research will focus on addressing these limitations and exploring the applicability of the approach in more complex settings and challenging applications.

Acknowledgments

This research was partially funded by the French National Research Agency (ANR) (Grant No. ANR-23-CPJ1-0099-01) and the France 2030 program under the reference 23-PEIA-004.

References

[1] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. Cambridge University Press, 2009.

[2] J. Pearl and D. Mackenzie, *The book of Why: the new Science of Cause and Effect*. Basic books, 2018.

[3] D. Jacob, "CATE Meets ML: conditional average treatment effect and machine learning," *Digital Finance*, vol. 3, no. 2, pp. 99–148, 2021.

[4] L. Nie, M. Ye, Q. Liu, and D. Nicolae, "VCNet and functional targeted regularization for learning causal effects of continuous treatments," in *International Conference on Learning Representations*, 2021.

[5] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[6] A. Menon and C. S. Ong, "Linking losses for density ratio and class-probability estimation," in *International Conference on Machine Learning*, 2016, pp. 304–313.

[7] K. Hirano and G. W. Imbens, "The propensity score with continuous treatments," *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, vol. 226164, pp. 73–84, 2004.

[8] X. Wang, S. Lyu, X. Wu, T. Wu, and H. Chen, "Generalization bounds for estimating causal effects of continuous treatments," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8605–8617, 2022.

Sharp Bounds for Continuous-Valued Treatment Effects with Unobserved Confounders

Jean-Baptiste Baitairian [1] & Bernard Sebastien [2] & Rana Jreich [3] & Sandrine Katsahian [4] & Agathe Guilloux [5]

[1] *Sanofi & Inria, Université Paris Cité, Inserm, HeKA, France, jean-baptiste.baitairian@inria.fr*
[2] *Sanofi, France, bernard.sebastien@sanofi.com*
[3] *Sanofi, France, rana.jreich@sanofi.com*
[4] *Inria, Université Paris Cité, Inserm, HeKA & CIC-EC 1418 - Paris HEGP, France, sandrine.katsahian@aphp.fr*
[5] *Inria, Université Paris Cité, Inserm, HeKA, France, agathe.guilloux@inria.fr*

Abstract. In causal inference, treatment effects are typically estimated under the ignorability, or unconfoundedness, assumption, which is often unrealistic in observational data, as it posits no unobserved confounders. By relaxing this assumption and conducting a sensitivity analysis, we introduce novel bounds and derive confidence intervals for the Average Potential Outcome (APO) – a standard metric for evaluating continuous-valued treatment or exposure effects. We demonstrate that these bounds are sharp under a continuous sensitivity model, in the sense that they give the smallest possible interval under this model. In a comparative analysis with the method of Jesson et al. (2022), using both simulated and real datasets, we show that our approach not only yields sharper bounds but also achieves good coverage of the true APO, with significantly reduced computation times.

Keywords. Average Potential Outcome, causal inference, continuous-valued intervention, dose-response function, sensitivity analysis.

1 Introduction

Estimating the causal effects of continuous interventions is crucial across many domains, including life sciences or economics. In particular, observational data, or more generally real-world data, offer a significant opportunity to enhance clinical drug development and support regulatory decisions by using, for instance, Electronic Health Records (EHR), medical claims data or measurements from wearable devices. Since the 21st Century Cures Act, the United States Food and Drug Administration (FDA) has even been requested to develop a program in order to evaluate how real-world evidence could be used for medical product approvals (Ding et al., 2023).

To leverage observational data in the context of treatment effect estimation, most of the approaches rely on a strong assumption, known as *ignorability*, *unconfoundedness* or *exogeneity* (Rubin, 1974) in causal inference. This assumption posits that, inside a subgroup

of units that share similar observed characteristics, the treatment can be considered randomly assigned. In the case of continuous treatments, the *Average Potential Outcome* (APO), sometimes named *dose-response function*, stands out as a metric of reference to evaluate treatment effect. Under ignorability, noteworthy estimation approaches of the APO are *Inverse Probability Weighting*-like methods (Imai and Van Dyk, 2004; Hirano and Imbens, 2004; Kallus and Zhou, 2018; Colangelo and Lee, 2023). The conditional counterpart of the APO, the *Conditional Average Potential Outcome* (CAPO), can also be estimated via similar techniques.

Nevertheless, this hypothesis may be overly optimistic given the impossibility to observe all confounding variables, especially in observational studies. To overcome this issue, recent works have suggested using sensitivity models to bound the biased treatment effect estimates, providing intervals as solutions. In the binary treatment scenario, researchers have explored ways to relax the usual ignorability assumption using, for instance, the *Marginal Sensitivity Model* (MSM) (Tan, 2006; Zhao, Small, and Bhattacharya, 2019; Dorn and Guo, 2022; Dorn, Guo, and Kallus, 2024).

Following these works, Jesson et al. (2022) extended the MSM to the continuous case (*Continuous Marginal Sensitivity Model*, or CMSM) and provided bounds for the APO. Related works also include the ones from Bonvini et al. (2022) and Frauen, Melnychuk, and Feuerriegel (2024). In this paper, we present a new methodology that combines the doubly robust kernel-based APO estimator from Kallus and Zhou (2018) with an additional constraint on a weight function, and derive sharp bounds under a sensitivity model introduced in Section 2. Then, we deduce confidence intervals (CI) via the *percentile bootstrap method*, as done in Zhao, Small, and Bhattacharya (2019). Specifically, this novel approach provides tighter and faster-to-compute bounds for the APO, as compared to Jesson et al. (2022).

2 Problem setting and notations

In the following, we consider the *Neyman-Rubin potential outcome framework* (Neyman, 1923; Rubin, 1974) adapted to continuous treatments. We denote by $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^{p_\mathbf{X}}$ the vector of observed confounders, with $p_\mathbf{X} \geq 1$, $T \in \mathcal{T} \subseteq \mathbb{R}$ the continuous treatment or exposition, and $Y(t) \in \mathcal{Y} \subseteq \mathbb{R}$ the potential outcome for a treatment value t. If Y is the observed outcome and $T = t$, then we assume that $Y = Y(t)$ (*consistency* and *non-interference*). Estimated quantities are denoted with a hat $\hat{\cdot}$, unless stated otherwise. Conditional expectations are equally written with or without a subscript, e.g. $\mathbb{E}_{\mathbf{X}=\mathbf{x}}[Y] \coloneqq \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$.

For a fixed treatment or exposition value $\tau \in \mathcal{T}$ and covariate vector $\mathbf{x} \in \mathcal{X}$, we are interested in estimating and bounding the APO $\theta(\tau)$ and CAPO $\theta(\tau, \mathbf{x})$, which are defined as $\theta(\tau) \coloneqq \mathbb{E}[Y(\tau)]$ and $\theta(\tau, \mathbf{x}) \coloneqq \mathbb{E}_{\mathbf{X}=\mathbf{x}}[Y(\tau)]$. They are commonly estimated under the ignorability assumption for continuous treatments (Hirano and Imbens, 2004): $\forall t \in \mathcal{T}$, $Y(t) \perp\!\!\!\perp T \mid \mathbf{X}$. For convenience, we refer to it as $\mathbf{X}$-*ignorability*. Under this hypothesis and consistency of the outcomes, notice that

$$\theta(\tau) = \mathbb{E}[\theta(\tau, \mathbf{X})] = \mathbb{E}[\mathbb{E}[Y(\tau)|\mathbf{X}]] = \mathbb{E}[\mathbb{E}_{\mathbf{X},T=\tau}[Y(T)]] = \mathbb{E}[\eta(\tau, \mathbf{X})] \tag{1}$$

where $\eta(t, \mathbf{x}) \coloneqq \mathbb{E}_{\mathbf{X}=\mathbf{x}, T=t}[Y]$ is a conditional expectation that could be estimated by regression. Kallus and Zhou (2018) and Colangelo and Lee (2023) proposed simple, *stabilized* and/or *doubly robust* versions of a kernel-based estimator of the APO. We focus here on the doubly robust estimator from Kallus and Zhou (2018). For a fixed treatment value $\tau \in \mathcal{T}$, and considering an i.i.d. observed sample $\mathcal{D} = \{(\mathbf{X}_i, T_i, Y_i)\}_{i=1}^n$ of size n, they express it as

$$\hat{\theta}_h(\tau) = \hat{\eta}(\tau) + \frac{1}{n} \sum_{i=1}^{n} \frac{K_h(T_i - \tau)}{\hat{f}(T = T_i | \mathbf{X} = \mathbf{X}_i)} \big(Y_i - \hat{\eta}(T_i, \mathbf{X}_i)\big), \tag{2}$$

where $\hat{\eta}(\tau) = \sum_{i=1}^n \hat{\eta}(\tau, \mathbf{X}_i)/n$. K_h is defined as $K_h(s) = K(s/h)/h$, where $h > 0$ is a *bandwidth* and K is a *kernel*. Common choices of K include the Epanechnikov or Gaussian kernels. The purpose of K_h is to localize the estimation around the treatment of interest τ. Then, the density of T conditionally on $\mathbf{X} = \mathbf{x}$, $t \mapsto f(T = t | \mathbf{X} = \mathbf{x})$, is known as *Generalized Propensity Score* (GPS) (Hirano and Imbens, 2004). We assume that it exists and that it is positive for all $\mathbf{x} \in \mathcal{X}$ and $t \in \mathcal{T}$ (referred to as *positivity* assumption).

However, as mentioned earlier, $\mathbf{X}$-ignorability is rarely satisfied in practice due to the presence of unobserved confounders $\mathbf{U} \in \mathcal{U} \subseteq \mathbb{R}^{p_\mathbf{U}}$, where $p_\mathbf{U} \geq 1$. A more reasonable assumption is to impose $(\mathbf{X}, \mathbf{U})$-*ignorability*: $\forall t \in \mathcal{T}$, $Y(t) \perp\!\!\!\perp T | \mathbf{X}, \mathbf{U}$. It formalizes the idea that the unmeasured confounders $\mathbf{U}$ capture all common causes of T and $Y(t)$ not included in $\mathbf{X}$ (Kallus, Mao, and Uehara, 2021). Under this assumption, the final equality in (1) no longer holds, meaning that $\theta(\tau) \neq \mathbb{E}[\eta(\tau, \mathbf{X})]$, and the estimator of the APO from Equation (2) becomes biased.

In practice, to relax the $\mathbf{X}$-ignorability assumption, we work under the Continuous Marginal Sensitivity Model from Jesson et al. (2022).

Definition 2.1. (CMSM for the APO from Jesson et al. (2022)) Under positivity assumption, for a given treatment value $\tau \in \mathcal{T}$, for all $(\mathbf{x}, \mathbf{u}) \in \mathcal{X} \times \mathcal{U}$, there exists a *sensitivity parameter* $\Gamma \geq 1$ such that

$$\Gamma^{-1} \leq \frac{f(T = \tau | \mathbf{X} = \mathbf{x}, Y(\tau) = y)}{f(T = \tau | \mathbf{X} = \mathbf{x})} \leq \Gamma. \tag{3}$$

Note that Γ can be seen as a user-defined sensitivity parameter which measures the deviation from the usual $\mathbf{X}$-ignorability assumption: if $\Gamma = 1$, the CMSM becomes equivalent to $\mathbf{X}$-ignorability, as if all confounders were observed; higher values of Γ assume a greater effect of the unobserved confounders on the treatment T and a deviation from $\mathbf{X}$-ignorability.

3 Bounds for the CAPO and APO

Notice first that, by positivity assumption, the CAPO can be rewritten

$$\theta(\tau, \mathbf{x}) = \mathbb{E}_{\mathbf{X}=\mathbf{x}, T=\tau}[Y w^\star(Y, \mathbf{x}, \tau)],$$

where, for all $(\mathbf{x}, t, y) \in \mathcal{X} \times \mathcal{T} \times \mathcal{Y}$,

$$w^\star(y, \mathbf{x}, t) \coloneqq \frac{f(Y(t) = y | \mathbf{X} = \mathbf{x})}{f(Y = y | \mathbf{X} = \mathbf{x}, T = t)}$$

exists (Jesson et al., 2022). Note that the weight function, or likelihood ratio, $w^\star$ fulfills the constraint

$$\mathbb{E}_{\mathbf{X}=\mathbf{x},T=t}[w^\star(Y,\mathbf{x},t)] = 1 \tag{4}$$

for all $(\mathbf{x},t) \in \mathcal{X} \times \mathcal{T}$, and takes values within $[\Gamma^{-1},\Gamma]$.

Using the conditional expectation η and the previous constraint (4), the CAPO can also be written

$$\theta(\tau,\mathbf{x}) = \eta(\tau,\mathbf{x}) + \mathbb{E}_{\mathbf{X}=\mathbf{x},T=\tau}\big[\big(Y-\eta(\tau,\mathbf{x})\big)w^\star(Y,\mathbf{x},\tau)\big].$$

Under the CMSM (3), we therefore define the bounds for the CAPO as

$$\theta^-(\tau,\mathbf{x}) \coloneqq \eta(\tau,\mathbf{x}) + \inf_{w\in\mathcal{W}^\star_\tau} \mathbb{E}_{\mathbf{X}=\mathbf{x},T=\tau}\big[\big(Y-\eta(\tau,\mathbf{x})\big)w(Y,\mathbf{x},\tau)\big], \tag{5}$$

$$\theta^+(\tau,\mathbf{x}) \coloneqq \eta(\tau,\mathbf{x}) + \sup_{w\in\mathcal{W}^\star_\tau} \mathbb{E}_{\mathbf{X}=\mathbf{x},T=\tau}\big[\big(Y-\eta(\tau,\mathbf{x})\big)w(Y,\mathbf{x},\tau)\big], \tag{6}$$

where $\mathcal{W}^\star_\tau$ is the set of functions $w : \mathcal{Y}\times\mathcal{X}\times\mathcal{T} \to [\Gamma^{-1},\Gamma]$ that satisfy Equation (4). The solutions of the optimization problems (5) and (6) are given by the following theorem.

Theorem 3.1. *Under positivity assumption and the CMSM* (3)*, the solutions to the optimization problems* (5) *and* (6) *are given by*

$$\theta^\pm(\tau,\mathbf{x}) = \eta(\tau,\mathbf{x}) + \frac{2\gamma-1}{\gamma}\mathbb{E}_{\mathbf{X}=\mathbf{x},T=\tau}\Big[Y-\eta(\tau,\mathbf{x})\Big|Q^\pm_{\mathbf{X},\tau}\Big] \tag{7}$$

where $\pm$ *stands for either* $+$ *or* $-$*,* $Q^+_{\mathbf{X},\tau} \coloneqq \{Y > q^{\mathbf{X},\tau}_\gamma\}$*,* $Q^-_{\mathbf{X},\tau} \coloneqq \{Y \le q^{\mathbf{X},\tau}_{1-\gamma}\}$*,* $\gamma \coloneqq \Gamma/(1+\Gamma)$ *and* $q^{\mathbf{x},\tau}_\upsilon \coloneqq Q(\upsilon; Y|\mathbf{X}=\mathbf{x}, T=\tau)$*, the quantile of order* υ *of the distribution of* Y *conditionally on* $\mathbf{X}=\mathbf{x}$ *and* $T=\tau$*.*

Moreover, the partially identified set $[\theta^-(\tau,\mathbf{x}),\theta^+(\tau,\mathbf{x})]$ *is sharp under the CMSM* (3)*, in the sense that it is the smallest possible interval under the CMSM for a given sensitivity parameter* Γ*.*

Similarly, for the APO, we can show that the interval $[\theta^-(\tau),\, \theta^+(\tau)]$ is sharp under the CMSM (3), using the relation $\theta^\pm(\tau) = \mathbb{E}[\theta^\pm(\tau,\mathbf{X})]$.

As the treatment or exposition T is continuous, estimating the conditional expectations in Equation (7) requires a tool to localize the estimation around the treatment of interest τ. Nonparametric kernel regression provides a solution. As in Kallus and Zhou (2018), we use kernels and define below *kernelized* versions of the bounds for the APO that are indexed by a bandwidth $h > 0$:

$$\theta^\pm_h(\tau) \coloneqq \eta(\tau) + \frac{2\gamma-1}{\gamma}\,\mathbb{E}\left[\mathbb{E}\left[\frac{K_h(T-\tau)}{f(T|\mathbf{X})}(Y-\eta(\tau,\mathbf{X}))\middle|\mathbf{X}, Q^\pm_{\mathbf{X},T}\right]\right],$$

where K_h is defined as in Equation (2). Therefore, by defining $\mathcal{I}_{\mathcal{D}_-} = \{i \in [\![1,n]\!]\,;\, Y_i \le q^{\mathbf{X}_i,T_i}_{1-\gamma}\}$ of cardinality n_-, and $\mathcal{I}_{\mathcal{D}_+} = \{i \in [\![1,n]\!]\,;\, Y_i > q^{\mathbf{X}_i,T_i}_{\gamma}\}$ of cardinality n_+, we can obtain estimators of $\theta^-_h(\tau)$ and $\theta^+_h(\tau)$ as follows:

$$\tilde{\theta}^\pm_h(\tau) = \tilde{\eta}(\tau) + \frac{2\gamma-1}{\gamma n_\pm}\sum_{i\in\mathcal{I}_{\mathcal{D}_\pm}}\frac{K_h(T_i-\tau)}{f(T_i|\mathbf{X}_i)}(Y_i-\eta(T_i,\mathbf{X}_i))$$

where $\tilde{\eta}(\tau) = \frac{1}{n}\sum_{i=1}^{n}\eta(\tau, \mathbf{X}_i)$.

Finally, we follow Zhao, Small, and Bhattacharya (2019), Dorn and Guo (2022) and Jesson et al. (2022), and build a $(1-\alpha)$-confidence interval for the upper and lower bounds of the APO under our CMSM via the *percentile bootstrap method*.

4 Experiments

We perform different experiments on simulated and real data and show that our method provides sharp bounds and outperforms the existing methodology from Jesson et al. (2022) in terms of computation times and sharpness. Our R implementation is publicly available at https://github.com/Sanofi-Public/sharp-bounds-for-continuous-treatment-effects, under a non-commercial license. In Figure 1, we illustrate the behavior of our method on a publicly available dataset from the U.S. Environmental Protection Agency that studies the impact of annual PM2.5 fine particulate matter level on cardiovascular mortality rate (CMR) in 2,132 U.S. counties between 1990 and 2010 (Wyatt et al., 2020). As in Bahadori, Tchetgen, and Heckerman (2022), we restrict the data to year 2010 to simplify the study and we consider 10 continuous variables as observed confounders $\mathbf{X}$. Finally, as in the case of our simulated dataset, we remove 10% of extreme observations which leads to an effective sample size of $n = 1918$. Note that our bounds are still sharper than the ones from Jesson et al. (2022). Moreover, estimating Γ via *informal benchmarking* (Cinelli and Hazlett, 2020) on the 10 continuous covariates leads to a value greater than 3. Therefore, this sensitivity analysis cannot completely justify an effect of PM2.5 on CMR, as the CIs for $\Gamma = 3$ (between PM2.5 equals to 3.27 and 7.98 $\mu g/m^3$) contain the average CMR, i.e. the no effect horizontal curve.

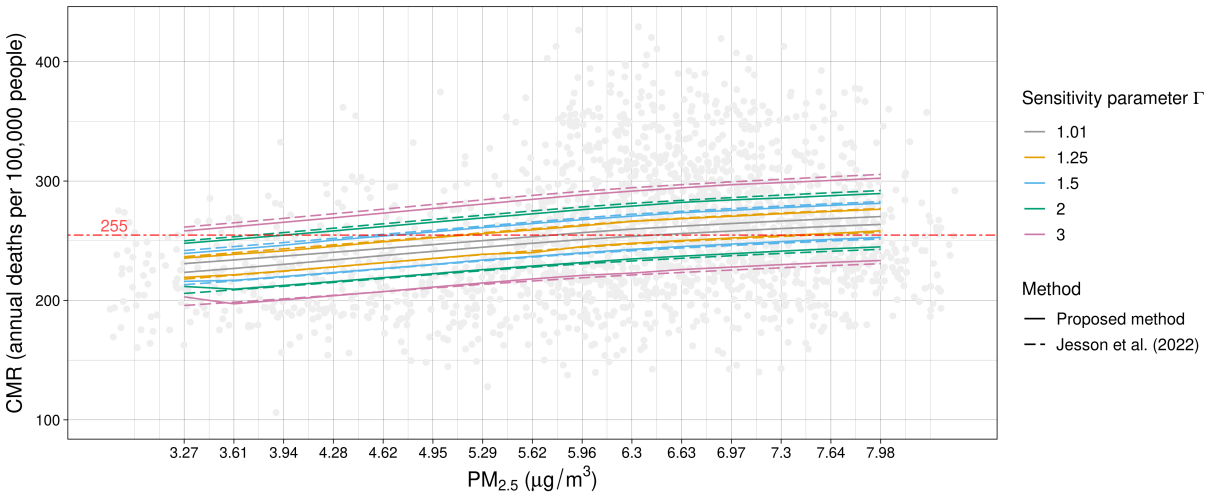


Figure 1: Sensitivity analysis (95%-level confidence intervals) with the proposed method and the one from Jesson et al. (2022). The red dotted line corresponds to the average CMR (255 annual deaths per 100,000 people). The gray points are the real dataset with 84 observations removed to improve readability.

5 Conclusion and disclosures

We presented novel bounds for the APO after introducing a new continuous sensitivity model and leveraging a constraint on a likelihood ratio, derived confidence intervals from them, and showed that our algorithm outperforms existing methodologies in terms of computation times and sharpness. These results were demonstrated on a simulated dataset and real data from the U.S. Environmental Protection Agency.

Jean-Baptiste Baitairian, Bernard Sebastien and Rana Jreich are Sanofi employees and may hold shares and/or stock options in the company. Agathe Guilloux is employed by INRIA. Sandrine Katsahian is employed by Université Paris-Cité and Assistance Publique – Hôpitaux de Paris (AP-HP). This work was supported by Sanofi and INRIA.

References

Bahadori, Taha, Eric Tchetgen Tchetgen, and David Heckerman (2022). "End-to-end balancing for causal continuous treatment-effect estimation". In: *International Conference on Machine Learning*. PMLR, pp. 1313–1326.

Bonvini, Matteo et al. (2022). "Sensitivity analysis for marginal structural models". In: *arXiv preprint arXiv:2210.04681*.

Cinelli, Carlos and Chad Hazlett (2020). "Making sense of sensitivity: Extending omitted variable bias". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.1, pp. 39–67.

Colangelo, Kyle and Ying-Ying Lee (2023). *Double Debiased Machine Learning Nonparametric Inference with Continuous Treatments*. arXiv: 2004.03036 [econ.EM]. URL: https://arxiv.org/abs/2004.03036.

Ding, Peng et al. (2023). "Sensitivity Analysis for Unmeasured Confounding in Medical Product Development and Evaluation Using Real World Evidence". In: *arXiv preprint arXiv:2307.07442*.

Dorn, Jacob and Kevin Guo (2022). "Sharp sensitivity analysis for inverse propensity weighting via quantile balancing". In: *Journal of the American Statistical Association*, pp. 1–13.

Dorn, Jacob, Kevin Guo, and Nathan Kallus (2024). "Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding". In: *Journal of the American Statistical Association*, pp. 1–12.

Frauen, Dennis, Valentyn Melnychuk, and Stefan Feuerriegel (2024). "Sharp bounds for generalized causal sensitivity analysis". In: *Advances in Neural Information Processing Systems* 36.

Hirano, Keisuke and Guido W Imbens (2004). "The propensity score with continuous treatments". In: *Applied Bayesian modeling and causal inference from incomplete-data perspectives* 226164, pp. 73–84.

Imai, Kosuke and David A Van Dyk (2004). "Causal inference with general treatment regimes: Generalizing the propensity score". In: *Journal of the American Statistical Association* 99.467, pp. 854–866.

Jesson, Andrew et al. (2022). "Scalable sensitivity and uncertainty analyses for causal-effect estimates of continuous-valued interventions". In: *Advances in Neural Information Processing Systems* 35, pp. 13892–13907.

Kallus, Nathan, Xiaojie Mao, and Masatoshi Uehara (2021). "Causal inference under unmeasured confounding with negative controls: A minimax learning approach". In: *arXiv preprint arXiv:2103.14029*.

Kallus, Nathan and Angela Zhou (2018). "Policy evaluation and optimization with continuous treatments". In: *International conference on artificial intelligence and statistics*. PMLR, pp. 1243–1251.

Neyman, Jerzy (1923). "On the application of probability theory to agricultural experiments. Essay on principles". In: *Ann. Agricultural Sciences*, pp. 1–51.

Rubin, Donald B (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5, p. 688.

Tan, Zhiqiang (2006). "A distributional approach for causal inference using propensity scores". In: *Journal of the American Statistical Association* 101.476, pp. 1619–1637.

Wyatt, Lauren H et al. (2020). "Annual PM2. 5 and cardiovascular mortality rate data: Trends modified by county socioeconomic status in 2,132 US counties". In: *Data in brief* 30, p. 105318.

Zhao, Qingyuan, Dylan S Small, and Bhaswar B Bhattacharya (2019). "Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 81.4, pp. 735–761.

Policy learning under constraint: seeking primary outcome maximization under adverse outcome control

Fuentes Vicente Laura [1] & Mathieu Even [2] & Julie Josse [3] & Antoine Chambaz [4]

[1] *Inria Premedical Inria-Idesp (Inserm - University of Montpellier) team, France, laura.fuentes-vicente@inria.fr*
[2] *Inria Premedical Inria-Idesp (Inserm - University of Montpellier) team, France, mathieu.even@inria.fr*
[3] *Inria Premedical Inria-Idesp (Inserm - University of Montpellier) team, France, julie.josse@inria.fr*
[4] *MAP5 Paris Cité University, France, antoine.chambaz@u-paris.fr*

Résumé. Traditionnellement, une politique assigne un traitement aux patients en fonction de leurs caractéristiques afin de maximiser la réponse attendue, en supposant que le traitement est administré conformément à cette recommandation. Une politique est dite optimale si elle maximise un ou plusieurs critères.

Une approche courante consiste à attribuer un traitement en fonction du signe de l'effet moyen conditionnel du traitement (Conditional Average Treatment Effect, CATE), une mesure causale qui quantifie l'effet du traitement au sein d'un sous-groupe spécifique de la population. Si le CATE est positif, le traitement est considéré comme bénéfique, tandis que s'il est négatif, il peut être inefficace ou même nocif. Cette approche repose donc sur l'apprentissage d'un estimateur précis du CATE. Cependant, bien qu'efficace dans certains contextes, elle présente des limites dans les cadres à résultats multiples et ne prend pas en compte les effets indésirables, un élément pourtant essentiel à considérer dans la prescription de traitements en milieu clinique.

Le défi que nous abordons est d'apprendre un estimateur optimal du CATE tout en garantissant que la politique qui en découle respecte les contraintes liées aux effets indésirables. Nous nous concentrons sur les situations où la probabilité d'événements indésirables augmente lorsque le traitement est administré.

Pour répondre à cette problématique, nous proposons une approche s'inscrivant dans le cadre classique de l'optimisation sous contrainte. Notre méthode équilibre simultanément la précision de l'apprentissage du CATE et le respect des contraintes imposées à la politique dérivée. Nous y parvenons en intégrant un paramètre d'échelle, un multiplicateur de Lagrange, directement dans la fonction objective à minimiser.

Mots-clés. Inférence causale, Règles de traitement individualisées, optimisation de politiques, optimisation sous contraintes.

Abstract. Traditionally, a policy maps patient characteristics to a treatment recommendation with the aim of maximizing the expected response, assuming treatment is allocated according to the policy's recommendation. It is considered optimal if it maximizes one or more criteria.

A policy may assign treatment based on the sign of the Conditional Average Treatment Effect (CATE), a causal measure that quantifies the treatment effect within a specific subgroup of a population. If the CATE is positive, then the treatment has a beneficial effect. Otherwise, the treatment may be harmful or ineffective. This approach shifts the problem toward learning an accurate estimator of the CATE. However, while this strategy works well in certain contexts, it is not properly suited for multiple-outcome frameworks and overlooks adverse events, which are essential to consider when prescribing treatments in a clinical setting.

The challenge we address is learning an estimator for the CATE in such a way that the resulting policy adheres to constraints related to adverse outcomes. Our focus is on cases where the probability of adverse events increases when treatment is administered. We propose an approach that fits within the classical constrained optimization framework. Our method simultaneously accounts for the risk of learning the CATE accurately and the constraint on the derived policy. This is achieved by incorporating a scaling parameter, a Lagrangian multiplier, into the objective function that we aim to minimize.

Keywords. Causal inference, Individualized treatment rules, Policy optimization, Constrained optimization

1 Causal setting

Let $\mathcal{O}_1, \ldots, \mathcal{O}_i = (X_i, T_i, Y_i, Z_i), \ldots, \mathcal{O}_n$ be n independent observations drawn from an unknown distribution P_0. For each $i \in [n]$, $X_i \in \mathcal{X}$ contains the i-th patient's characteristics, $T_i \in \mathcal{T} = \{0, 1\}$ indicates which of two treatments was administered, $Y_i \in [0, 1]$ is their primary outcome, measuring how well the treatment worked (with convention the higher, the better), and $Z_i \in \mathbb{R}^p$ is their secondary outcome, measuring physiological features related to adverse effects.

We adopt the potential outcomes framework (Rubin, 2005), and further introduce complete data-structures that encompass potential outcomes, formalizing what would have occurred in two worlds where each possible treatment would have been administered to every patient. This yields the n complete data-structures

$$\mathbb{O}_i = (X_i, T_i, Y_i(1), Y_i(0), Z_i(1), Z_i(0), Y_i, Z_i), \quad i \in [n]$$

drawn independently from an unknown distribution $\mathbb{P}_0$ that encompasses P_0.

1.1 The CATE and its identifiability

The Conditional Average Treatment Effect (CATE):

$$\Delta_{\mathbb{P}_0}(X) = E_{\mathbb{P}_0}[Y(1) - Y(0) \mid X] \tag{1}$$

quantifies the treatment effect within a subgroup characterized by covariates $X \in \mathcal{X}$. A fundamental challenge in causal inference is to establish assumptions that enable the identifiability of $\Delta_{\mathbb{P}_0}(X)$ (a feature of $\mathbb{P}_0$) under P_0 (Imbens and Rubin, 2015). It is well known that if *stable unit treatment-value* (SUTVA), *positivity* and *conditional ignorability* hold (Rubin, 1980; Rosenbaum and Rubin, 1983), then the CATE is identifiable. Specifically, denoting $\mu_{P_0,t}(X) = E_{P_0}[Y \mid X, T = t]$ for both $t \in \mathcal{T}$,

$$\Delta_{\mathbb{P}_0}(X) \stackrel{\triangle}{=} \mu_{P_0,1}(X) - \mu_{P_0,0}(X) = \Delta\mu_{P_0}(X). \tag{2}$$

1.2 Policy learning

An essential question in personalized medicine is *which treatment works best for whom?* When only two treatments are available, determining the optimal treatment assignment for each patient can be thought as a classification task.

A policy π is an element of a user-supplied set $\Pi \subset [0,1]^{\mathcal{X}}$. It maps patient covariates to a treatment recommendation. If the mapping is maximized with respect to one or several criteria, then the policy is said optimal.

Value-driven optimality. The value of a policy $\pi \in \Pi$ taking its values in $\mathcal{T}$, defined as

$$\mathcal{V}_{\mathbb{P}_0}(\pi) = E_{\mathbb{P}_0}[Y(\pi(X))], \tag{3}$$

is the average of the outcome in a world where the policy's recommendations are followed. It is a natural criterion which we seek to maximize. If the assumptions of Section 1.1, hold, maximizing (3) is equivalent to maximizing

$$E_{P_0}[\pi(X)\Delta\mu_{P_0}(X)], \tag{4}$$

which is a feature of P_0. We now denote by $\pi^*(P_0)$ a maximizer of (4). In addition, we emphasize that (4) is meaningful for any $\pi \in \Pi$ beyond the policies that only take the 0 or 1.

Learning a value-optimal policy. Learning a value-optimal policy can be performed directly or indirectly.

The direct approach consists of performing a classification task in the treatment space $\mathcal{T}$, aiming to predict the sign of $\Delta\mu_{P_0}(X)$. This can be done by single-stage outcome weighted learning (Montoya et al., 2023) and by weighted classification (Wager, 2022), for example. Alternatively, because

$$x \mapsto \mathbb{1}\{\Delta\mu_{P_0}(x) > 0\} \tag{5}$$

maximizes (4), the indirect approach focuses on the estimation of $\Delta\mu_{P_0}$. Traditional plug-in estimators (Wager, 2022) rely heavily on the accuracy of outcome regression models, while risk minimization approaches aim for robustness by leveraging nuisance parameters (Kennedy, 2023). The latter often employ Neyman orthogonal losses to achieve efficiency, but they come with challenges such as non-convex optimization and extreme estimates of CATE.

In van der Laan et al. (2024), the authors introduce the EP-learner, a novel approach based on an efficient plug-in (EP) risk estimator for directly estimating $\Delta\mu_{P_0}$. It targets the projection of $\Delta\mu_{P_0}$ onto the user-supplied convex set $\Psi \subset [-1,1]^{\mathcal{X}}$, that is the minimizer over Ψ of the risk

$$\psi \mapsto R_{P_0}(\psi) = E_{P_0}\left[\psi(X)^2 - 2\psi(X)\Delta\mu_{P_0}(X)\right].$$

This helps preserving the stability and robustness of plug-in estimators while achieving asymptotic equivalence to the oracle-efficient one-step debiased estimator under mild assumptions.

In view of (5), an estimator of $\Delta\mu_{P_0}$ easily maps to a policy. Instead of proceeding that way, we *smoothly* map every $\psi \in \Psi$ to a policy

$$\pi = \sigma_\beta \circ \psi \in \Pi = \{\sigma_\beta \circ \psi' : \psi' \in \Psi\}, \quad \text{with}$$
$$\sigma_\beta(u) \propto \log\left(\frac{1+e^{\beta u}}{1+e^{-\beta}}\right) \quad (\text{all } u \in [-1,1]).$$

We interpret $\pi(x) = \sigma_\beta \circ \psi(x)$ as the probability of recommending treatment $t = 1$ conditionally on $X = x$, where β controls the scaling of $\Delta\mu_{P_0}$ to the $[0,1]$ range. This resolves both differentiability issues associated with a rule as in (5) and provides a measure of confidence in treatment recommendation.

However, such an approach does not account for potential adverse effects Z, which are often critical when prescribing treatments.

2 Policy learning under constraint

Several studies have explored the incorporation of various constraints into risk minimization tasks (Nabi et al., 2024), as well as in policy optimization problems. The constraints may concern fairness (Fang et al., 2023), limited resources (Luedtke and van der Laan, 2016), the trade-off between maximizing benefits and controlling risks (Wang et al., 2018), among others.

In Section 2.1, we reframe our question of interest in terms of policy learning under constraint. In Section 2.2, we discuss how to tackle the problem empirically and we present the results of a simulation study.

2.1 Oracular viewpoint

Formalization. Let $\mathcal{Z}$ be a set of undesirable outcomes. Our goal is to control the probability that a patient experiences the adverse event $Z \in \mathcal{Z}$.

Assumption 2.1. *We assume that treatment has a deleterious effect on the adverse event,*

$$\mathbb{P}_0(Z(1) \in \mathcal{Z}|X) \geq \mathbb{P}_0(Z(0) \in \mathcal{Z}|X).$$

Under SUTVA, positivity and conditional ignorability assumptions, this is equivalent to

$$\Delta\nu_{P_0}(X) = P_0(Z \in \mathcal{Z}|T = 1, X) - P_0(Z \in \mathcal{Z}|T = 0, X) \geq 0.$$

We wish to guarantee that the policies we consider do not increase, in average, the probability of the adverse event beyond a user-supplied $\alpha \in [0, 1/2]$, using $\mathbb{P}_0(Z(0) \in \mathcal{Z}|X)$ as a baseline. To do so, we introduce the constraint

$$\psi \mapsto S_{P_0}(\psi) = E_{P_0}[\sigma_\beta \circ \psi(X)\Delta\nu_{P_0}(X)]] - \alpha$$

which maps Ψ to $\mathbb{R}$. Under SUTVA, positivity and conditional ignorability assumptions, $\alpha + S_{P_0}(\psi)$ can be interpreted as the average (with respect to the marginal distribution of X under $\mathbb{P}_0$) of the difference between $\mathbb{P}_0(Z(0) \in \mathcal{Z}|X)$ (viewed as a baseline) and the counterfactual conditional probability that a patient described by X experience the adverse event in a world where the policy $\sigma_\beta \circ \psi$ would be systematically applied.

Now, the policy learning under constraint optimization problem can be simply formulated as

$$\begin{aligned} &\text{minimize } R_{P_0}(\psi) \text{ with respect to } \psi \in \Psi \\ &\text{subject to } S_{P_0}(\psi) \leq 0. \end{aligned} \tag{6}$$

On solving (6)**.** This type of problem can be solved using a Lagrangian. Specifically, the Lagrangian can be defined here as the mapping from $\Psi \times \mathbb{R}_+$ to $\mathbb{R}$ given by

$$(\psi, \lambda) \mapsto R_{P_0}(\psi) + \lambda S_{P_0}(\psi) = U_{P_0}(\psi, \lambda).$$

This formulation enables a trade-off between optimizing treatment outcomes and controlling adverse effects.

2.2 Empirical viewpoint and simulations

We face several challenges, presented below as questions. The first set of questions concern the oracular U_{P_0}.

1. Does $(\psi, \lambda) \mapsto U_{P_0}(\psi, \lambda)$ admit critical points?
2. If so, does there exist a closed-form expression for the critical λ (as a feature of P_0)? If not, can the critical λ be (sharply?) upper and lower-bounded by quantities known in closed-form (as features of P_0)?

Naturally, the oracular U_{P_0} is unknown and we need to estimate it. The second set of questions concern the estimation of U_{P_0} with U_n and the properties of the critical points of U_n. We will rely on a cross-fitting procedure to estimate the key-nuisance parameters $\Delta\mu_{P_0}$ and $\Delta\nu_{P_0}$.

1. What theoretical guarantees can we give regarding the estimation of $R_{P_0}(\psi)$ and $S_{P_0}(\psi)$?
2. What theoretical guarantees can we give regarding the estimation of the critical points of U_n?

We will implement the estimation procedure. It will yield the estimated rule $\pi_n = \sigma_\beta \circ \psi_n$.

3. What is the value of the rule π_n?
4. How does it compare with those of competing rules derived by following other estimation strategies?

We will present the results of a simulation study on both synthetically generated and real datasets.

References

Fang, E. X., Wang, Z., and Wang, L. (2023). Fairness-oriented learning for optimal individualized treatment rules. *Journal of the American Statistical Association*, 118(543):1733–1746.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049.

Luedtke, A. R. and van der Laan, M. J. (2016). Optimal individualized treatments in resource-limited settings. *The international journal of biostatistics*, 12(1):283–303.

Montoya, L. M., van der Laan, M. J., Luedtke, A. R., Skeem, J. L., Coyle, J. R., and Petersen, M. L. (2023). The optimal dynamic treatment rule superlearner: considerations, performance, and application to criminal justice interventions. *The international journal of biostatistics*, 19(1):217–238.

Nabi, R., Hejazi, N. S., van der Laan, M. J., and Benkeser, D. (2024). Statistical learning for constrained functional parameters in infinite-dimensional models with applications in fair machine learning.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

van der Laan, L., Carone, M., and Luedtke, A. (2024). Combining t-learning and dr-learning: a framework for oracle-efficient estimation of causal contrasts. *arXiv preprint arXiv:2402.01972*.

Wager, S. (2022). Stats 361: Causal inference.

Wang, Y., Fu, H., and Zeng, D. (2018). Learning optimal personalized treatment rules in consideration of benefit and risk: with an application to treating type 2 diabetes patients with insulin therapies. *Journal of the American Statistical Association*, 113(521):1–13.

Measuring Variable Importance in Heterogeneous Treatment Effects with Confidence

Joseph Paillard[1,2], Angel Reyero Lobo[3], Vitaliy Kolodyazhniy[1], Bertrand Thirion[2,4], & Denis A. Engemann[1]

[1]*Roche Pharma Research Early Development, Basel, Switzerland, joseph.paillard@roche.com*
[2] *Inria, Universite Paris Saclay, France* [3] *Institut de Mathematiques de Toulouse ; UMR5219 Université de Toulouse, France* [4]*CEA*

Résumé. Le machine learning causal est prometteur pour estimer les effets individuels des traitements à partir de données complexes. Pour garantir le succès d'applications utilisant ces méthodes, il est essentiel d'obtenir des informations fiables sur les variables qui entraînent une hétérogénéité dans la réponse au traitement. Nous proposons permucate, un algorithme basé sur la méthode Conditional Permutation Importance (CPI), pour mesurer l'importance globale des variables, avec des garanties statistiques, pour l'estimation du CATE. L'analyse théorique non-asymptotique et les études empiriques montrent que permucate a une variance plus faible que la méthode de référence Leave-One-Covariate-Out (LOCO) et fournit une mesure fiable de l'importance des variables. Cette propriété augmente la puissance statistique, ce qui est crucial pour l'inférence causale dans un domaine dans lequel peu de données sont disponibles, typiquement pour des applications au domaine biomédical. Nous démontrons empiriquement les avantages de permucate dans des ensembles de données de santé simulés et réels, y compris des contextes avec jusqu'à des centaines de variables corrélées.

Mots-clés. Inférence causale, importance de variables, machine learning ...

Abstract. Causal machine learning (ML) holds promise for estimating individual treatment effects from complex data. For successful real-world applications using machine learning methods, it is of paramount importance to obtain reliable insights into which variables drive heterogeneity in the response to treatment. We propose `PermuCATE`, an algorithm based on the Conditional Permutation Importance (CPI) method, for statistically rigorous global variable importance assessment in the estimation of the Conditional Average Treatment Effect (CATE). Theoretical analysis of the finite sample regime and empirical studies show that `PermuCATE` has lower variance than the Leave-One-Covariate-Out (LOCO) reference method and provides a reliable measure of variable importance. This property increases statistical power, which is crucial for causal inference in the limited-data regime common to biomedical applications. We empirically demonstrate the benefits of `PermuCATE` in simulated and real-world health datasets, including settings with up to hundreds of correlated variables.

Keywords. Causal inference, variable importance, machine learning ...

1 Introduction

The ever growing volume of high-dimensional biomedical data holds promise to facilitate the discovery of novel biomarkers that can predict individual treatment success or risk of adverse events [15]. Transforming complex biological signals into actionable variables, therefore, plays a crucial role for increasing the efficiency of biomedical research and development [6]. Machine Learning (ML) promises to improve the prediction of biomedical and clinical outcomes from high-dimensional, high-throughput data in various biomedical applications due to its ability to learn complex functions and representations from heterogeneous and correlated data, such as those encountered in e.g. imaging, genomics or transcriptomics data [11]. Regardless, their notorious lack of interpretability hinders the applicability of machine-learning models for generating scientific and clinical insights, motivating efforts for new methods for estimating variable importance within the field of interpretable ML [14]. Most studies so far have focused on classical ML models like random forests [17], which are well-suited for tabular data with few variables [7]. A more recent line of work, exposed to complex life-science data, has therefore emphasised the importance of high-capacity ML algorithms and proposed methods combining high-capacity ML algorithms with inference algorithms backed by statistical guarantees. [13, 19].

Despite significant efforts and advances on methods for interpretability, the utility of ML models remains limited as biomedical data are often of interventional nature. Therefore, such datasets are typically analyzed using biostatistical methods appropriate for causal inference and accepted by regulatory agencies. Over the past decade, findings from the causal inference and potential outcomes literature [9] have been adapted to propose the estimand framework, a flexible statistical-modelling framework useful for causal analysis of clinical data. It provides formal concepts and draws a conceptual distinction between the quantity of interest and the estimation method. This provides a formal language that can bridge the gap between classical biostatistics and ML, hence, pave the way for causal machine learning in biomedical applications [5, 4]. However, when studying heterogeneity of treatment effects using machine-learning, explainability with statistical guarantees is crucial for uncovering useful predictive biomarkers as ad-hoc methodology may increase the false positives rate as well as the false negative rate.

Taken together, in the context of present and future biomedical applications, ML methods must address both the complexity of multimodal high-dimensional biological data and support the research questions that arise from interventional studies with estimands such as average treatment effects (ATE) and conditional average treatment effects (CATE). In the causal setting, it remains unclear whether variable importance can be measured with intact statistical guarantees.

Contributions. In this work we study algorithms for estimating the importance of variables with statistical guarantees in the context of interventional data and CATE.

- We propose an new variable importance inference procedure for heterogeneous treatment effects termed `PermuCATE`, which generalizes Conditional Permutation Importance (`CPI`) [2]. We put it in perspective with the established leave-one-covariate-out (`LOCO`) method [19, 12].

- We develop a theoretical analysis of the finite sample variance for both `PermuCATE` and `LOCO`, focusing on the impact of estimation error in finite samples.
- Through experiments on simulated and real-world data, using various causal ML models, we show how `PermuCATE` achieves lower variance and higher statistical power due to the influence of finite sample estimation error, an effect that is more pronounced in small sample sizes or when using complex deep learning CATE estimators.

2 `PermuCATE` algorithm

The method presented in this section aims at identifying important variable for CATE prediction. While similarly to LOCO, it is agnostic to the meta-learning approach used to estimated the CATE, we used the DR-learner [10]. This choice is motivated by its favorable convergence properties and the fact that the scoring step requires deriving pseudo-outcomes explicitly, which limits the computational benefits of simpler and more direct meta-learning approaches such as the S- or T-learner.

Algorithm 1 Conditional Permutation Importance for CATE

Input: D_{train}, D_{test} two independent sets of observations $Z_i = (X_i, A_i, Y_i), X_i \in R^d$
Parameter: Response functions: $\hat{\mu}_0, \hat{\mu}_1$; Propensity: $\hat{\pi}$; CATE: $\hat{\tau}$; Covariate predictor: $\hat{\nu}(\cdot)$, feasible risk: $R(\cdot)$, Number of permutations: P
Output: Importance estimate $\hat{\Psi} \in R^d$

Using D_{train}
Fit $\hat{\mu}_0, \hat{\mu}_1, \hat{\pi}, \hat{\tau}$ // CV can be used for DR-learner
for $j = 1, \cdots, d$ **do**
    Fit $\hat{\nu}_j(\cdot) \leftarrow \hat{E}_n[X^j | X^{-j}]$
end for
Using D_{test}
for $j = 1, \cdots, d$ **do**
    $\hat{\nu}_j \leftarrow \hat{\nu}_j(X^{-j})$
    $r_j \leftarrow X^j - \hat{\nu}_j$
    for $k = 1, \cdots, P$ **do**
        $\widetilde{r_j} \leftarrow \text{shuffle}(r_j)$
        $X_{P,j} \leftarrow [X_1, \cdots, X_{j-1}, \hat{\nu}_j + \widetilde{r_j}, \cdots, X_d]$
        $\hat{\Psi}^j_k \leftarrow (R(\hat{\tau}(X_{P,j}), Y, A) - R(\hat{\tau}(X), Y, A))/2$
    end for
    $\hat{\Psi}^j \leftarrow \frac{1}{P} \sum_{k=1}^{P} \hat{\Psi}^j_k$
end for
return $[\hat{\Psi}^1, \cdots, \hat{\Psi}^d]$

`LOCO` estimates the importance of a given variable by refitting a model on a subset of covariates, whereas permutation-based approaches reuse the same model to predict from a perturbed design matrix. In the traditional permutation importance approach breiman2001random,

the perturbed matrix differs from the original by a permutation of the studied covariate. That approach is naive in the sense that it does not account for correlation, leading to uncontrolled type-1 errors; `CPI`, a refined version mitigates this issue [2]. The key idea is to sample from the conditional distribution $p(X^j|X^{(-j)})$ which requires the following assumption.

Assumption 1: There exists a function ν_j such that $X^j = \nu_j(X^{(-j)}) + \varepsilon_j$ with $\varepsilon_j \perp\!\!\!\perp X^{(-j)}$ and $E[\varepsilon_j] = 0$.

Sampling from the conditional can thus be achieved by estimating ν_j with a consistent estimator $\hat{\nu}_j$ (see Proposition 3.3 from [16]). Then, the residual information $r_j = X^j - \hat{\nu}_j$ can used to construct a perturbed version of the j^{th} covariate: $\hat{\nu}_j + \widetilde{r_j}$ where $\widetilde{r_j} = \text{shuffle}(r_j)$ is a permutation of the residual.

Proposition 1: Assume the consistency of the estimators $\hat{\tau}$ and assumption:1. `PermuCATE` consistently estimates $2E\left[V\left[\tau(X)|X^{(-j)}\right]\right]$, twice the quantity known as the total Sobol index. The proof is provided in appendix.

Remark 1: assumption:1 amounts to assuming that the relationships between the covariates within X are simpler than the link between the CATE, τ and the input X. Consequently, estimating the conditional density is simpler than estimating the outcome, akin to the model-X knockoff framework presented in [1]. Numerous examples illustrate the utility of this setting, such as in the study of biologically complex diseases using multiple variables like single nucleotide polymorphisms (SNPs) where the relationship between covariates is much simpler than their association with clinical outcomes. Similarly, in clinical trials, the covariates are collected concurrently (e.g. at baseline) but the outcome of interest is measured at a later time. Furthermore, the amount of unlabelled data often exceeds the number of available labels, especially in clinical settings where labelling can be costly. Adapting this idea to the context of CATE estimation, given a causal risk R, the following measure of variable importance can be obtained:

$$\Psi^j_{CPI} = \frac{1}{2}\left(R\left(\tau, X_{P,j}, A, Y\right) - R\left(\tau, X, A, Y\right)\right) \tag{1}$$

where $X_{P,j} = [X_1, \cdots, X_{j-1}, \nu_j + \widetilde{r_j}, \cdots, X_d]$. `PermuCATE` (Algorithm 1) describes the procedure to obtain Ψ^j_{CPI}. Previous works focusing on predictive models have shown how a cross-fitting approach allows to obtain a valid estimate of the desired quantity with a control of the type-I error [19, 18, 2, 16].

Properties The following paragraphs compare the properties of the `LOCO` and `PermuCATE` estimators in the context of finite training sets $\mathcal{D}_{\text{train}}$, to highlight variance-inducing error terms from finite sample estimations. A key advantage of `PermuCATE` is that it allows the use of an expressive model for the main function estimation, such as a deep neural network, without incurring the high cost for refitting the full model, provided that a more parsimonious model is used for covariate prediction. However, in finite sample regimes, this estimation of $\hat{\nu}_j$ on a fixed train set D_{train} introduces variance.

Proposition 2 Under *Assumption 1* on $\hat{\nu}_j$, for a consistent and Lipschitz CATE estimator $\hat{\tau}$, the finite sample estimation of the importance for the j^{th} covariate is

$$\frac{1}{2}\widehat{\Psi}^j_{\text{CPI}}|\mathcal{D}_{\text{train}} - \Psi^*_j = O_P(\|\nu_j - \widehat{\nu}_j\|^2) \tag{2}$$

where $\Psi_j^* = E\left[V\left[\tau(X)|X^{(-j)}\right]\right]$ is the targeted total Sobol index. And $||\cdot||^2$ is the l2-norm given the fixed training set, $\mathcal{D}_{\text{train}}$.
Proof in appendix.
In line with proposiiton 1, it can be observed that, given a consistent estimator of ν_j, $\widehat{\Psi}^j_{\text{CPI}}$ is a consistent estimator of the total Sobol index. It also shows that the error comes from estimating $\hat{\nu}_j$ To better understand the impact of estimation error on variance, consider a linear setting with Gaussian covariates where the CATE is defined as $\tau(X) = \beta X$, with a finite training sample and an infinitely large testing sample, the variance of the `PermuCATE` estimator is given by

$$\text{var}\left(\frac{1}{2}\widehat{\Psi}^j_{\text{CPI}}|\mathcal{D}_{\text{train}}\right) = \text{var}(\widehat{\beta}_j^2||\Delta\widehat{\gamma}||^2_{\Sigma_{-j}}) \tag{3}$$

where $\Delta\widehat{\gamma} = \gamma - \widehat{\gamma}$ is the error when estimating the coefficients of the function $\nu_j(X^{-j}) = \gamma X^{-j}$ on the training set, $\mathcal{D}_{\text{train}}$ and Σ_{-j} is the covariance matrix of X^{-j} (proof in appendix). Consequently, (3) demonstrates that the variance of `PermuCATE` is mainly influenced by the error in estimating the conditional distribution via $\hat{\nu}$ but not on the error in $\hat{\tau}$. Concerning the `LOCO` approach, the iterative refitting of the model on the subsets of covariates X^{-j} leads to fluctuations due to the estimation of $\hat{\tau}$ and introduces a bias term associated with the CATE prediction.
Propoition 3 For a consistent CATE estimator $\hat{\tau}$, the finite sample estimation error of the importance for the j^{th} covariate is

$$\widehat{\Psi}^j_{\text{LOCO}}|\mathcal{D}_{\text{train}} - \Psi_j^* = O_P\left(\|\tau_{-j} - \widehat{\tau}_{-j}\|^2 - \|\tau - \widehat{\tau}\|^2\right)$$

Proof in appendix.
Denoting $\Delta\hat{\beta}$ and $\Delta\hat{\beta}_{-j}$ the difference between the estimated and true coefficients for respectively $\tau(X) = \beta X$ and the sub-model τ_{-j}. In the linear scenario as presented above the finite sample variance of the `LOCO` estimator is then

$$\text{var}\left(E\left[\widehat{\Psi}^j_{\text{LOCO}}\middle|\mathcal{D}_{\text{train}}\right]\right) = \text{var}\left(||\Delta\hat{\beta}_{(-j)}||^2_{\Sigma_{-j}}\right) + \text{var}\left(||\Delta\hat{\beta}||^2_{\Sigma}\right)$$

In the above equations, which present the different bias and variances in a linear setting assuming independent training samples, a key distinction emerges between the two approaches. The variance of `LOCO` depends on the sum of the error in estimating the CATE $\hat{\tau}$ and its sub-model $\hat{\tau}_j$. In contrast, for `PermuCATE` it depends on the error in estimating the conditional distrubution (through the error in $\hat{\nu}_j$). As discussed in remark:1, the latter prediction problem is arguably easier to solve.

Therefore, we investigated the variance behavior of both approaches using finite-sample simulations and datasets. The results in the following section will show empirically that the error term in `LOCO` can lead to increased variance and reduced statistical power.

3 Experiments

Figure 1, left, `PermuCATE` **detects important variables with greater statistical power.** Using the data-generating process proposed by [8], we compared the estimates of variable

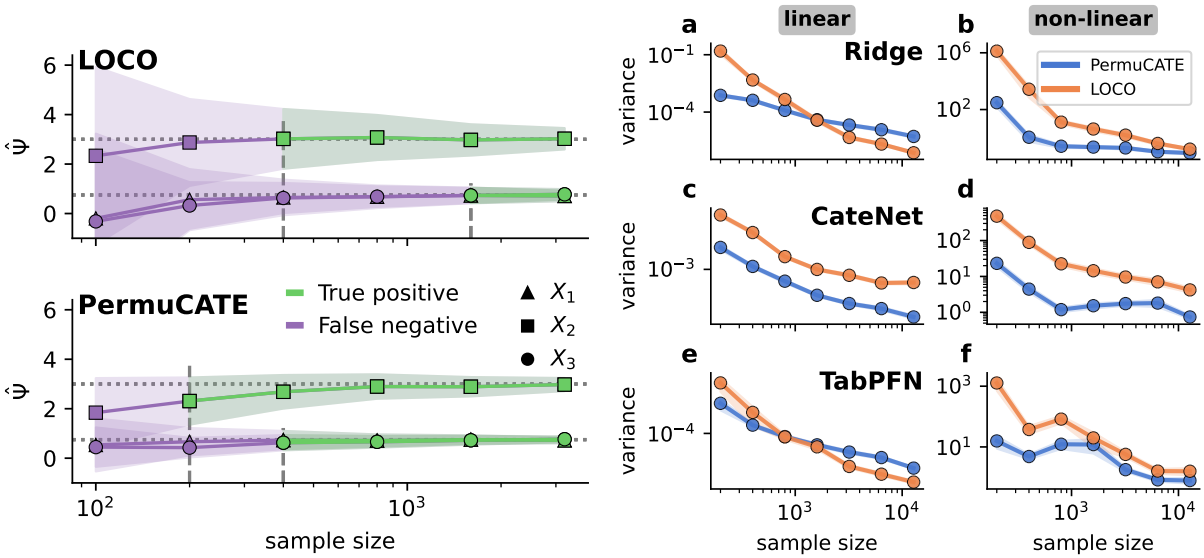


Figure 1

importance using `LOCO` (top subplot), the baseline proposed by the authors, and `PermuCATE`. By computing p-values, we observed at different sample sizes whether each of the three important variables was correctly identified (true positive) or missed (false negative) at a significance level $\alpha = 0.05$. For all three important variables, the minimum sample size (vertical dashed line) required to detect their importance was smaller with the `PermuCATE` approach. Both methods converged to the theoretical importance score (horizontal dotted line). The shaded areas around the curves represent Monte Carlo estimates of the standard deviation over 100 random samples. Figure 1, right, **PermuCATE demonstrates lower variance for misspecified models and small sample sizes.** The figure compares the variance of the CATE estimations when the true CATE is a linear function of the covariates (left column) and a polynomial function of degree three with interaction terms (right column). Three different CATE estimators were used: (**a, b**) DR-learner with Ridge to estimate the nuisances and pseudo-outcomes; (**c, d**) CateNet from [3]; and (**e, f**) DR-learner using TabPFN. Both simulation scenarios used (d=50) correlated Gaussian covariates with a support size of 10. Variance was assessed for one predictive variable across 100 random simulations. Error bars represent bootstrap estimates of the variance.

References

[1] Emmanuel Candes et al. "Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.3 (2018), pp. 551–577.

[2] Ahmad Chamma, Denis A Engemann, and Bertrand Thirion. "Statistically Valid Variable Importance Assessment through Conditional Permutations". In: *Advances in Neural Information Processing Systems* (2023). DOI: 10.48550/arxiv.2309.07593.

[3] Alicia Curth and Mihaela van der Schaar. "Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms". In: *Advances in Neural Information Processing Systems* (2021). DOI: 10.48550/arxiv.2101.10943.

[4] Matthieu Doutreligne and Gaël Varoquaux. "How to select predictive models for causal inference?" In: *arXiv* (2023). DOI: 10.48550/arxiv.2302.00370.

[5] Stefan Feuerriegel et al. "Causal machine learning for predicting treatment outcomes". In: *Nature Medicine* 30.4 (2024), pp. 958–968.

[6] Richard Frank and Richard Hargreaves. "Clinical biomarkers in drug discovery and development". In: *Nature reviews Drug discovery* 2.7 (2003), pp. 566–580.

[7] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on typical tabular data?" In: *Advances in neural information processing systems* 35 (2022), pp. 507–520.

[8] Oliver Hines, Karla Diaz-Ordaz, and Stijn Vansteelandt. "Variable importance measures for heterogeneous causal effects". In: *arXiv* (2022). DOI: 10.48550/arxiv.2204.06030.

[9] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[10] Edward H Kennedy. "Towards optimal doubly robust estimation of heterogeneous causal effects". In: *Electronic Journal of Statistics* (2023). DOI: 10.1214/23-ejs2157.

[11] Alan Le Goallec et al. "Using deep learning to predict abdominal age from liver and pancreas magnetic resonance images". In: *Nature Communications* 13.1 (2022), p. 1979.

[12] Jing Lei et al. "Distribution-free predictive inference for regression". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1094–1111.

[13] Xinlei Mi et al. "Permutation-based identification of important biomarkers for complex diseases via machine learning models". In: *Nature communications* 12.1 (2021), p. 3008.

[14] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[15] Menglan Pang et al. "Precision medicine analysis of heterogeneity in individual-level treatment response to amyloid beta removal in early Alzheimer's disease". In: *Alzheimer's & Dementia* (2023). DOI: 10.1002/alz.13431.

[16] Angel Reyero Lobo, Pierre Neuvial, and Bertrand Thirion. "Sobol-CPI: a Doubly Robust Conditional Permutation Importance Statistic". In: (2025). URL: https://arxiv.org/abs/2501.17520.

[17] Carolin Strobl et al. "Conditional variable importance for random forests". In: *BMC bioinformatics* 9 (2008), pp. 1–11.

[18] Isabella Verdinelli and Larry Wasserman. "Feature Importance: A Closer Look at Shapley Values and LOCO". In: *arXiv* (2023). DOI: 10.48550/arxiv.2303.05981.

[19] Brian D. Williamson et al. "A General Framework for Inference on Algorithm-Agnostic Variable Importance". In: *Journal of the American Statistical Association* (2023). DOI: 10.1080/01621459.2021.2003200.

Choix de variable en Survie Causale

Charlotte Voinot [1] & Julie Josse [2] & Bernard Sebastien [3]

[1] *INRIA Premedical - SANOFI, France, charlotte.voinot@sanofi.com*
[2] *INRIA, France, julie.josse@inria.fr*
[3] *SANOFI, France, bernard.sebastien@sanofi.com*

Résumé. L'analyse de survie causale combine l'analyse de survie et l'inférence causale pour évaluer l'effet d'un traitement sur un temps de survie. Différentes approches, basées sur la pondération, la régression ou des méthodes double/triple robustes, permettent d'estimer l'effet moyen du traitement (Average Treatment Effect, ATE), mesuré via la différence entre les temps de survie moyen restreint (Restricted Mean Survival Time, RMST). Ces estimations peuvent être obtenues à partir de données issues d'essais randomisés ou de cohortes observationnelles, dans des contextes de censure indépendante ou dépendante.

Ces estimateurs reposent sur des modèles distincts pour le traitement, la réponse (le temps de survie) et la censure. Toutefois, les variables utilisées dans ces modèles ne sont généralement pas différenciées en fonction de leur rôle. L'objectif de ce travail est d'étudier l'impact de l'inclusion de variables additionnelles sur la variance des estimateurs. Nous distinguons en particulier trois types de variables : les variables de précision, qui influencent uniquement la réponse ; les variables instrumentales, qui affectent exclusivement le traitement ; et les variables spécifiques à la censure, qui n'impactent que le mécanisme de censure.

Dans la continuité des résultats classiques en inférence, nous montrons que l'intégration des variables de précision dans les modèles de réponse et de traitement réduit la variance des estimateurs. À l'inverse, l'ajout de variables exclusivement liées à la censure dans le modèle de censure tend à accroître cette variance. Ces résultats permettent de formuler des recommandations pratiques pour l'application des méthodes d'estimation en analyse de survie causale.

Mots-clés. Analyse de survie, Inférence causale, Survie causale, Restricted Mean Survival Time, Variance, Ensemble d'ajustement minimum, Ensemble d'ajustement étendu.

Abstract. Causal survival analysis combines survival analysis and causal inference to assess the effect of a treatment on survival time. Different approaches, based on weighting, regression, or doubly/triply robust methods, allow for the estimation of the Average Treatment Effect (ATE), measured through the difference in Restricted Mean Survival Time (RMST). These estimates can be obtained from randomized trials or observational cohorts, in contexts of independent or dependent censoring.

These estimators rely on distinct models for treatment, outcome (survival time), and censoring. However, the variables used in these models are generally not differentiated according tog to their role. The objective of this work is to study the impact of including additional variables on the variance of the estimators. In particular, we distinguish three types of variables: precision variables, which influence only the outcome; instrumental variables, which affect only the treatment; and censoring-specific variables, which impact only the censoring mechanism.

Building on classical inference results, we show that incorporating precision variables into the outcome and treatment models reduces the variance of the estimators. Conversely, adding variables exclusively related to censoring in the censoring model tends to increase this variance. These findings provide practical recommendations for the application of estimation methods in causal survival analysis.

Keywords. Survival analysis, causal inference, Causal survival, RMST, Variance, Minimal adjustment set, Extended adjustment set.

1 Context and Motivations

1.1 Causal inference with survival outcome

Notations and context Let $A \in \mathcal{A} = \{0, 1\}$ denote the binary treatment, $T \in \mathbb{R}^+$ the survival time, $C \in \mathbb{R}^+$ the censoring time, $\Delta = \mathbb{I}\{T \leq C\} \in \{0, 1\}$ the censoring indicator, and $\widetilde{T} = \min(T, C) \in \mathbb{R}^+$ the observed time. We adopt the potential outcomes framework, where for $a \in \mathcal{A}$, $T(a)$ represents the survival time if the subject had received treatment a (Rubin, 1974). Contrary to classical causal inference framework, $T(a)$ is not fully observed due to censoring. Let X be a p-dimensional vector of covariates belonging to a covariate space $X \subset \mathbb{R}^p$. We observe a sample of n independent and identically distributed (i.i.d.) observations $(X_i, A_i, \Delta_i, \widetilde{T}_i)_{i=1,\dots,n}$.

Figure 1 represents the causal graph associated to the problem in an observational setting with dependent censoring where the set X has been decomposed as follows: $X_{a,t}$ represents the vector of covariates that influence both A (treatment) and T (survival), i.e., confounders. Similarly, $X_{c,t}$ denotes the covariates affecting both C (censoring) and T (survival). The vector X_a includes covariates that influence only A (instrumental variables), while X_t consists of covariates affecting only T (precision variables). Lastly, X_c represents covariates that influence only C (censoring variables).

Quantity of interest The Average Treatment Effect (ATE) in causal survival analysis, can be expressed as the difference in Restricted Mean Survival Time (RMST):

$$\theta = \mathbb{E}\left[T(1) \wedge \tau - T(0) \wedge \tau\right].$$

where $\tau \in \mathbb{R}^+$ is the horizon time.

Identifiability Assumptions In observational studies with dependent censoring, the restricted mean survival time (RMST) can be identified from observed data under appropriate assumptions. These assumptions ensure that the treatment effect can be consistently estimated despite the challenges posed by confounding and informative censoring. We write these assumptions with the different set of features whereas in the literature, such assumptions are always written with the same X.

1. *Stable Unit Treatment Value Assumption (SUTVA)*: $T = AT(1) + (1 - A)T(0)$;

2. *Unconfoundedness*: $A \perp\!\!\!\perp \{T(1), T(0)\} \mid X_{a,t}$;

3. *Positivity for treatment*: $0 < \mathbb{P}(A = 1 \mid X_{a,t} = x) < 1$, for all $x \in \mathcal{X}_{a,t}$;

4. *Conditional independent censoring*: $C \perp\!\!\!\perp \{T(1), T(0)\} \mid X_{c,t}, A$;

5. *Positivity for censoring*: $\mathbb{P}(C \geq t \mid X_{c,t} = x, A = a) > 0$, for all $x \in \mathcal{X}_{c,t}$.

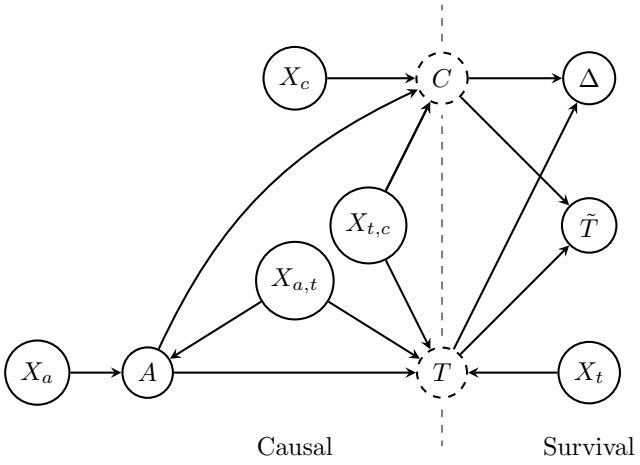


Figure 1: Causal graph for observational study with dependent censoring with different disjoint set of covariates

Estimation of the RMST Voinot et al. (2025) have reviewed the different estimators available for the RMST, namely the G-formula (Chen and Tsiatis, 2001), IPTW-Buckley-James (IPTW-BJ) (Buckley and James, 1979), IPTW-IPCW Kaplan-Meier (KM) (Schaubel and Wei, 2011), and the doubly/triply robust AIPTW-AIPCW (Ozenne et al., 2020), and have detailed their statistical properties (using only the set X). These estimators are based on the estimation of intermediary quantities that are called nuisance parameters. These include $e(x) = \mathbb{E}[A \mid X = x]$ the propensity score, $G(t|x, a), a \in \{0, 1\}$ the left-limit of the conditional survival function of the censoring $\mathbb{P}(C \geq t|X = x, A = a)$, $\mu(x, a) := \mathbb{E}[T \wedge \tau \mid X = x, A = a]$, $a \in \{0, 1\}$ the conditional restricted mean survival time and $S(t|x, a), a \in \{0, 1\}$ and the conditional survival function, $\mathbb{P}(T > t|X = x, A = a)$. Table 1 summarizes the nuisance parameters required for each of the RMST estimator.

1.2 Related work

In the broader field of causal inference, the role of precision and instrumental variables in treatment effect estimation has been examined: Instruments, should be removed from the adjustment set as this could increase variance (Zhou and Jia, 2021; Schnitzer et al., 2015)

Estimator	Outcome model	Censoring model	Treatment model
IPCW-IPTW-KM		G	e
IPTW-BJ	S		e
G-formula	μ		
AIPTW-AIPCW	S, μ	G	e

Table 1: Estimators of the difference in RMST and nuisance parameters needed to compute each estimator. Empty boxes indicate that the nuisance parameter is not needed in the estimator.

and precision covariates include in all estimators whether weighting methods (Lunceford and Davidian, 2004), regression (White and Lu, 2011) and methods combining both approaches.

Despite these findings, the impact of these variables in the context of causal survival analysis remains unexplored. Similarly, covariates that affect only censoring have been largely overlooked in treatment effect estimation.

This article aims to bridge the gap between causal inference and causal survival analysis by extending previous work on precision variables (X_t) and instrumental variables (X_a), while introducing new insights into the role of censoring variables (X_c).

We will examine how incorporating $X_a, X_t, X_c, X_{t,c}$ into nuisance parameters can affect estimators variance, providing a comprehensive analysis of their impact on bias and efficiency in survival settings.

2 Simulation results

Table 2 presents the combinations of sets tested in our simulations. The minimal sufficient set corresponds to the smallest set of covariates required to accurately identify the causal effect of the treatment (Neal, 2010).

Estimator	Outcome model	Censoring model	Treatment model
Minimal sufficient set	$T \sim X_{a,t}$	$C \sim X_{t,c}$	$A \sim X_{a,t}$
Extended set with X_t	$T \sim X_{a,t} + X_t$	$C \sim X_{t,c} + X_t$	$A \sim X_{a,t} + X_t$
Extended set with X_a	$T \sim X_{a,t} + X_a$	$C \sim X_{t,c} + X_a$	$A \sim X_{a,t} + X_a$
Extended set with X_c	$T \sim X_{a,t} + X_c$	$C \sim X_{t,c} + X_c$	$A \sim X_{a,t} + X_c$
Other unexplored Extended set	$T \sim X_{a,t} + X_{t,c}$ $T \sim X_{a,t} + X_{t,c} + X_t$		

Table 2: Different adjustement set for causal survival analysis. Red models correspond to the combination of set tested in the Figure 2

Figure 2 is representative of the results observed in different scenarios of simulation and highlighs the impact of different sets of variables on G-formula, IPTW-BJ and AIPTW-AIPCW estimators. The green boxplots correspond to the minimal set. The corresponding simulation generates an observational study with conditionally independent censoring. We

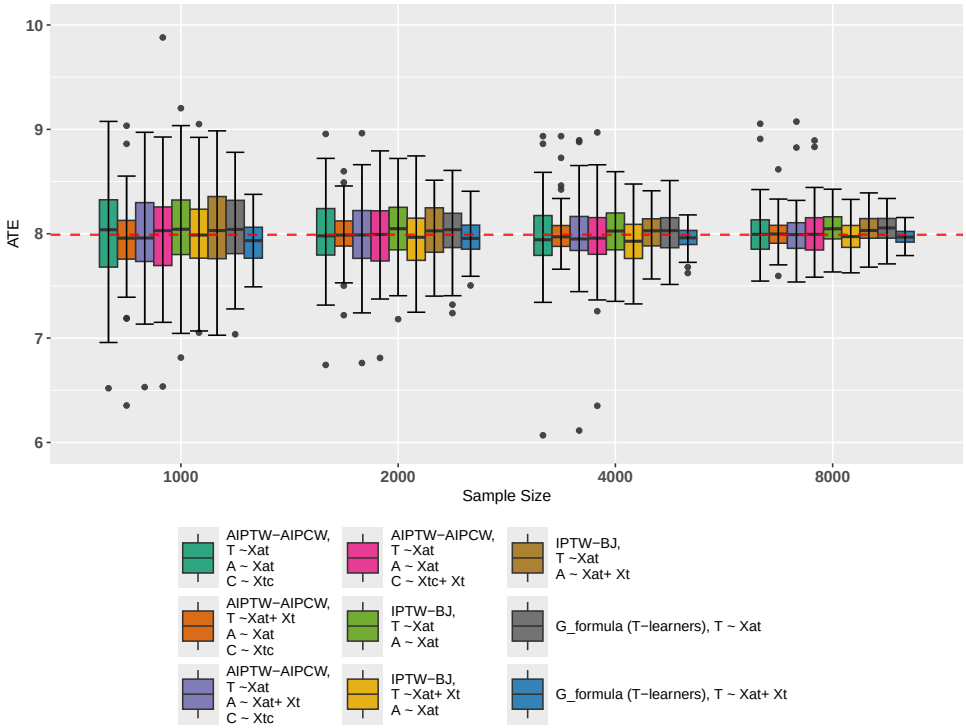


Figure 2: Average Treatment Effect estimation with different estimators and different sets of variables used within each estimator in the setting of an observational study with dependent censoring (The threshold time τ is set to 15).

sample n iid datapoints $(X_i, A_i, C, T_i(0), T_i(1))_{i\in[n]}$ where $T_i(0)$, $T_i(1)$, and C follow Cox's models. More specifically, we set:

- **Covariates:** $X_t \sim \mathcal{N}(\mu_t, \Sigma)$, $X_{a,t} \sim \mathcal{N}(\mu_{a,t}, \Sigma)$, $X_{t,c} \sim \mathcal{N}(\mu_{t,c}, \Sigma)$, where $\Sigma = \mathrm{Id}_2$, $\mu_t = (1, 1)$, $\mu_{a,t} = (0.5, 1)$, and $\mu_{t,c} = (0.5, 1)$.

- **The propensity score:** $\mathrm{logit}(e(X_{a,t})) = \beta_A^\top X_{a,t}$, where $\beta_A = (0.3, 1.1)$, where $\mathrm{logit}(p) = \log(p/(1-p))$.

- **The hazard function of $T(0)$:** $\lambda^{(0)}(t|X_t, X_{a,t}, X_{t,c}) = 0.01 \exp\left\{\beta_{t,0}^\top X_t + \beta_{at,0}^\top X_{a,t} + \beta_{tc,0}^\top X_{t,c}\right\}$, where $\beta_{t,0} = (2, 2)$, $\beta_{at,0} = (0.1, 0.1)$, and $\beta_{tc,0} = (0.5, 0.1)$.

- **The survival times in the treatment group:** $T(1) = T(0) + 10$.

- **The hazard function of the censoring time C:** $\lambda_C(t|X_{t,c}) = 0.015 \cdot \exp\left\{\beta_C^\top X_{t,c}\right\}$, where $\beta_C = (1, 1)$.

Adding precision variables (X_t) to the outcome model T sharply reduces the variance of the AIPTW-AIPCW and G-formula estimators. In contrast, its impact on the IPTW-BJ estimator is less pronounced.

Regarding the treatment model A, the inclusion of precision variables (X_t) does not have

a straightforward effect. It appears to slightly reduce the variance of the AIPTW-AIPCW and IPTW-BJ estimators, but this requires confirmation through theoretical results. In another simulation, we observe that the inclusion of instrumental variables (X_a) increases the variance of the AIPTW-AIPCW and IPTW-BJ estimator.

As for the censoring model C, incorporating variables related only to censoring (X_c) slightly increases the variance of the AIPTW-AIPCW estimator. We also observed that variables affecting both censoring and outcome ($X_{t,c}$) can improve precision when included in the outcome model. These results emphasize the importance of a careful variable selection for estimator performance in causal survival analysis. We will derive the associated variance to give the theoritical counterpart of these empirical results.

References

Buckley and James (1979, 12). Linear regression with censored data. Biometrika 66(3), 429–436.

Chen, P.-Y. and A. A. Tsiatis (2001). Causal inference on the difference of the restricted mean lifetime between two groups. Biometrics 57(4), 1030–1038.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine 23(19), 2937–2960.

Neal, B. (2010). Introduction to causal inference. J. Mach. Learn. Res. 11, 1643–1662.

Ozenne, B. M. H., T. H. Scheike, L. Stærk, and T. A. Gerds (2020). On the estimation of average treatment effects with right-censored time to event outcome and competing risks. Biometrical Journal 62(3), 751–763.

Rubin, D. B. (1974, October). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66(5), 688–701.

Schaubel, D. E. and G. Wei (2011, 03). Double Inverse-Weighted Estimation of Cumulative Treatment Effects under Nonproportional Hazards and Dependent Censoring. Biometrics 67(1), 29–38.

Schnitzer, M., J. Lok, and S. Gruber (2015, 07). Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. The international journal of biostatistics 12.

Voinot, C., C. Berenfeld, I. Mayer, B. Sebastien, and J. Josse (2025). Causal survival analysis, estimation of the average treatment effect (ate): Practical recommendations.

White, H. and X. Lu (2011, 11). Causal diagrams for treatment effect estimation with application to efficient covariate selection. The Review of Economics and Statistics 93(4), 1453–1459.

Zhou, K. and J. Jia (2021). Variance reduction for causal inference.

Informed learning

Conditionnement d'un processus gaussien avec des contraintes en inégalité par projection sur une famille de B-splines

Paul Castéras [1,2,3] & Gwenaël Salin [1]

[1] *CEA, DAM, DIF, F-91297, Arpajon, France, paul.casteras@cea.fr, gwenael.salin@cea.f*

[2] *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France*

[3] *CMAP, CNRS, École polytechnique, Institut polytechnique de Paris, 91120 Palaiseau, France*

Résumé. La régression par processus gaussien est une méthode de métamodélisation couramment employée pour remplacer un code de simulation coûteux. Nous disposons parfois d'informations supplémentaires sur les sorties du code. Par exemple, les sorties peuvent être positives, croissantes ou convexes. Utiliser un métamodèle qui prend en compte ces informations permet d'améliorer les prédictions. Une méthode pour intégrer ces contraintes au métamodèle consiste à considérer un processus qui s'écrit comme la somme d'un nombre fini de fonctions. Les conditions à respecter pour que les contraintes soient vérifiées s'écrivent alors sous la forme d'un nombre fini d'inégalités. Nous allons ici choisir de travailler avec des familles de fonctions qui sont des B-splines. Nous verrons alors que nous pouvons choisir la régularité du processus en modifiant l'ordre des B-splines. Cependant lorsque la régularité du processus augmente, les conditions à respecter pour que les contraintes soient vérifiées sont plus complexes, ce qui augmente le coût de calcul de la méthode. Après avoir présenté cette méthode, nous l'appliquerons sur un jeu de données d'équations d'état.

Mots-clés. Métamodèle, Processus gaussiens, Contraintes en inégalité

Abstract. Gaussian process regression is a surrogate method widely used to replace costly simulation codes. Sometimes, we have additional information about the outputs of the code. For example, the outputs may be positive, increasing or convex. Using a surrogate model that takes this information into account can improve the predictions. To incorporate these constraints into the surrogate model, one method considers a process written as the sum of a finite number of functions. The conditions required to satisfy the constraints can then be written as a finite number of inequalities. Here, we will choose to work with families of functions that are B-splines. We will see that the smoothness of the process can be controlled by modifying the order of the B-splines. However, as the smoothness of the process increases, the conditions required to satisfy the constraints become more complex, which increases the computational cost of the method. After presenting this method, we will apply it to a dataset of equations of state.

Keywords. Surrogate models, Gaussian processes, Inequality constraints

1 Introduction

Les simulations multiphysiques demandent des temps de calcul importants. Pour étudier les incertitudes liées aux modèles physiques, il est cependant nécessaire d'appeler un nombre important de fois un code de simulation. Un métamodèle qui approche le code de simulation, à partir des sorties du code dont nous disposons, est donc nécessaire afin de prédire la sortie du code de simulation ainsi que la confiance associée à cette prédiction. Une méthode de métamodélisation largement utilisée est la régression par processus gaussien (voir par exemple (Williams and Rasmussen, 2006)).

En plus de connaître la sortie d'une simulation pour un nombre fini d'entrées, nous pouvons avoir à notre disposition des contraintes supplémentaires en fonction du phénomène physique étudié (positivité, croissance, convexité,...). Pour améliorer la qualité du métamodèle, il est souhaitable d'intégrer ces contraintes dans ce dernier. Ces contraintes doivent être vérifiées sur un domaine continu. Il n'est donc pas possible de conditionner directement le processus gaussien afin que ces contraintes soient vérifiées sur tout le domaine. Différentes méthodes ont été proposées pour ajouter des contraintes en inégalité à la méthode de la régression par processus gaussien. L'article (Swiler et al., 2020) présente et discute différentes méthodes pour répondre à ce problème. Une approche proposée par Da Veiga and Marrel (2012) consiste à imposer la contrainte en un nombre fini de points. Si la contrainte est vérifiée en un nombre assez grand de points, alors nous pouvons espérer que la contrainte soit vérifiée sur tout le domaine, au moins approximativement.

L'approche que l'on va considérer ici, proposée par Maatouk (2015), consiste à projeter le processus gaussien sur une famille finie de fonctions. Il est alors possible d'imposer une contrainte sur tout le domaine en imposant un nombre fini de contraintes linéaires. López-Lopera et al. (2018) expliquent comment imposer des contraintes de borne, de croissance et de convexité en projetant le processus sur une famille de fonctions linéaires par morceaux. La famille de fonction choisie impose une régularité au processus projeté. Maatouk et al. (2024) propose d'adapter les familles de fonctions utilisées par Maatouk (2015) afin d'avoir une famille de fonction plus régulière pour les contraintes sur la dérivée ou la dérivée seconde. Si nous acceptons de travailler avec un nombre fini de contraintes non linéaires, nous allons ici voir qu'il est possible de travailler avec les familles de fonctions utilisées par Maatouk (2015) et d'avoir un processus projeté plus régulier pour les différents types de contraintes.

Dans la partie suivante, nous allons expliquer comment conditionner un processus gaussien avec des contraintes en inégalité en le projetant sur une base de B-splines. Nous allons dans la partie 3 appliquer cette méthode à un problème d'estimation d'équations d'état.

2 Processus gaussien avec contraintes en inégalité dans une base de B-splines

Dans cette partie, nous nous plaçons dans un premier temps dans le cadre de données d'entrée en dimension 1. Les B-splines seront introduits puis les contraintes pour des fonctions qui

s'écrivent comme une somme de B-splines seront définies. Nous verrons ensuite comment projeter un processus gaussien sur une base de B-splines permet d'ajouter des contraintes dans le modèle. Enfin, nous expliquerons comment appliquer cette méthode en dimension 2.

2.1 Projection d'une fonction et des contraintes sur des B-splines

Les B-splines sont des fonctions polynomiales par morceaux, définies sur un ensemble de réels $t_0 < t_1 < ... < t_N$ que nous appellerons nœuds. Nous supposons ici que ces nœuds sont régulièrement espacés (*i.e.* pour $i = 2, ..., N$, $t_i - t_{i-1} = t_1 - t_0$). Les B-splines sont utilisées pour approcher une fonction dont nous connaissons la valeur en un nombre fini de points en prenant en compte des critères de régularité (par exemple fonction continue, de classe $\mathcal{C}^1$ ou $\mathcal{C}^2$) en fonction de la famille de B-spline choisie. Nous renvoyons à (Kunoth et al., 2018) pour une présentation détaillée.

Nous notons B_i^k le i-ème B-spline d'ordre k. Les fonctions B-splines d'ordre k sont de classe $\mathcal{C}^{k-1}$ et sont des polynômes d'ordre k par morceaux. Pour une fonction $s \in \mathcal{C}^{k-1}$, on définit l'approximation $s_c(x) = \sum_{i=0}^{N-k-1} c_i.B_i^k(x)$ de s où les $c_i \in \mathbb{R}$, sont définis de manière unique en imposant :

$$\text{pour } l = 0, ..., k-2, s^{(l)}(t_k) = s_c^{(l)}(t_k) \text{ et pour } j = k, ..., N-k, s^{(k-1)}(t_j) = s_c^{(k-1)}(t_j). \quad (1)$$

Nous notons $\mathcal{G}^k$ l'opérateur linéaire qui à $s \in \mathcal{C}^{k-1}$ associe l'unique $c \in \mathbb{R}^{N-k}$ vérifiant la condition 1. La fonction s_c est alors de classe $\mathcal{C}^{k-1}$, et est un polynôme d'ordre k par morceau. Cette approximation permet de ramener les contraintes sur s_c à un nombre fini de contraintes sur les coefficients c. De plus, on montre qu'il suffit de s'intéresser aux contraintes de la fonction en positivité, car les contraintes d'autres types (fonctions bornées, croissantes, convexes) s'y ramènent. Pour le cas $k = 1$, la positivité de s_c est équivalente à la positivité de chacun des coefficients c.

À partir des résultats énoncés de Schmidt and Heß (1988), nous pouvons obtenir le même type de résultat pour $k = 2$ et $k = 3$. Plus précisément, la condition de positivité sur s_c peut s'écrire avec un nombre fini d'inégalités polynomiales sur les composantes du vecteur c. Cependant, les inégalités obtenues ne sont plus linéaires mais polynomiales. Dans ce cas, puisque le coût calculatoire de l'algorithme d'échantillonnage que nous utilisons augmente, nous proposons d'approcher ces contraintes polynomiales par des contraintes linéaires. Nous pouvons alors introduire des contraintes trop strictes et similaires au cas $k = 1$ ($c \geq 0$) ou pas assez strictes et similaires à la méthode proposée par Da Veiga and Marrel (2012) (pour $i = k, .., N-k, s_c(t_i) \geq 0$). Notons que ces deux approximations tendent vers les contraintes polynomiales lorsque le nombre de nœuds tend vers l'infini.

2.2 Échantillonner le processus projeté

Supposons maintenant que nous avons un processus gaussien centré $Y \sim \mathcal{GP}(0, \Sigma(\cdot, \cdot))$ de noyau de covariance Σ ainsi que n observations bruitées $(x_1, y_1), ..., (x_n, y_n)$ avec un bruit de mesure gaussien.

Soit $k \in \mathbb{N}$, on pose $c = \mathcal{G}^k(Y)$ et $Y_c(x) = \sum_{i=0}^{N-k-1} c_i.B_i^k(x)$. Nous supposons des conditions de régularité sur Σ de sorte que c soit bien défini. Y_c est alors un processus projeté sur la base de B-splines qui approche le processus initial Y. Comme $\mathcal{G}^k$ est linéaire, c est un vecteur gaussien dont la loi dépend du noyau de covariance Σ. Nous souhaitons maintenant déterminer sa loi, ce qui permettra de déterminer la loi de Y_c.

Comme expliqué par López-Lopera et al. (2018), nous pouvons alors déterminer la loi de c conditionnellement aux observations et aux contraintes en inégalité. En effet, la loi de c conditionnellement aux observations est une loi normale dont on peut déterminer la moyenne et la matrice de covariance. Le conditionnement par rapport aux contraintes revient à considérer que c suit une loi gaussienne tronquée. Il est alors possible d'estimer la moyenne et la matrice de covariance du vecteur c (et donc de Y_c) par une méthode de Monte-Carlo par chaine de Markov (MCMC). Nous utiliserons alors l'algorithme proposé par Pakman and Paninski (2014) qui utilise l'algorithme Hamiltonien Monte Carlo (HMC) pour des lois gaussiennes tronquées avec des contraintes polynomiales.

2.3 Généralisation en dimension 2

Pour réaliser l'application qu'on se propose de traiter, il est nécessaire de généraliser en dimension 2 la méthode précédente. En définissant des B-splines en dimension 2 comme un produit de B-splines de dimension 1, nous pouvons généraliser en dimension 2 la définition de la fonction s_c, et de l'opérateur $\mathcal{G}^k$, et pour $k = 1$, nous avons toujours que s_c positive est équivalent à c positif. Par contre, pour $k = 2$ et $k = 3$, le coût calculatoire induit par les contraintes polynomiales rend impossible l'application de l'algorithme d'échantillonnage. Nous pouvons cependant toujours travailler avec les deux approximations linéaires des contraintes polynomiales qui se généralisent à la dimension 2.

3 Application à des équations d'états

Les simulations hydrodynamiques nécessitent la connaissance d'équations d'état des matériaux, qui représentent des relations entre les variables d'état d'un système. Plus précisément, on s'intéresse ici à la pression P et à l'énergie interne E, comme fonctions de la densité ρ, et de la température T d'un matériau. Une introduction aux équations d'état est proposée par Eliezer et al. (1986). L'application qu'on se propose de traiter concerne une table d'équation d'état du fer calculée à partir du modèle semi classique de Thomas Fermi. Nous introduisons l'énergie libre $F = E - TS$, où S est l'entropie car la connaissance de F comme fonction de ρ et T détermine entièrement l'équation d'état d'un matériau. En effet, $P = \rho^2 \frac{\partial F}{\partial \rho}$ et $E = F - T\frac{\partial F}{\partial T}$. Nous allons appliquer la méthode introduite précédemment car nous disposons des contraintes en inégalité suivantes, qui se ramènent à des contraintes sur les dérivées secondes de F : $\frac{\partial P}{\partial \rho} \geq 0$ et $\frac{\partial E}{\partial T} \geq 0$.

Après avoir adimensionné les quantités physiques ($f = \frac{F}{T}$, $p = \frac{P}{\rho T} = \frac{\partial f}{\partial log \rho}$, et $e = -\frac{E}{T} = \frac{\partial f}{\partial log T}$), nous allons appliquer la méthode présentée précédemment pour intégrer ces con-

traintes en inégalité avec $k = 3$, $N = 40$, un noyau de covariance gaussien et des contraintes linéaires pas assez strictes. Le processus qui approche f dépend alors de $(\log T, \log \rho)$ et est conditionné par rapport aux observations non bruitées sur f, ainsi que ses dérivées e et p aux points $(x_i)_{i=1,\dots,n}$ en plus des contraintes en inégalité. Cette méthode est appliquée pour des températures allant de 10^4 K à 10^6 K et pour une densité entre 7.8 g/cm^3 et 78 g/cm^3.

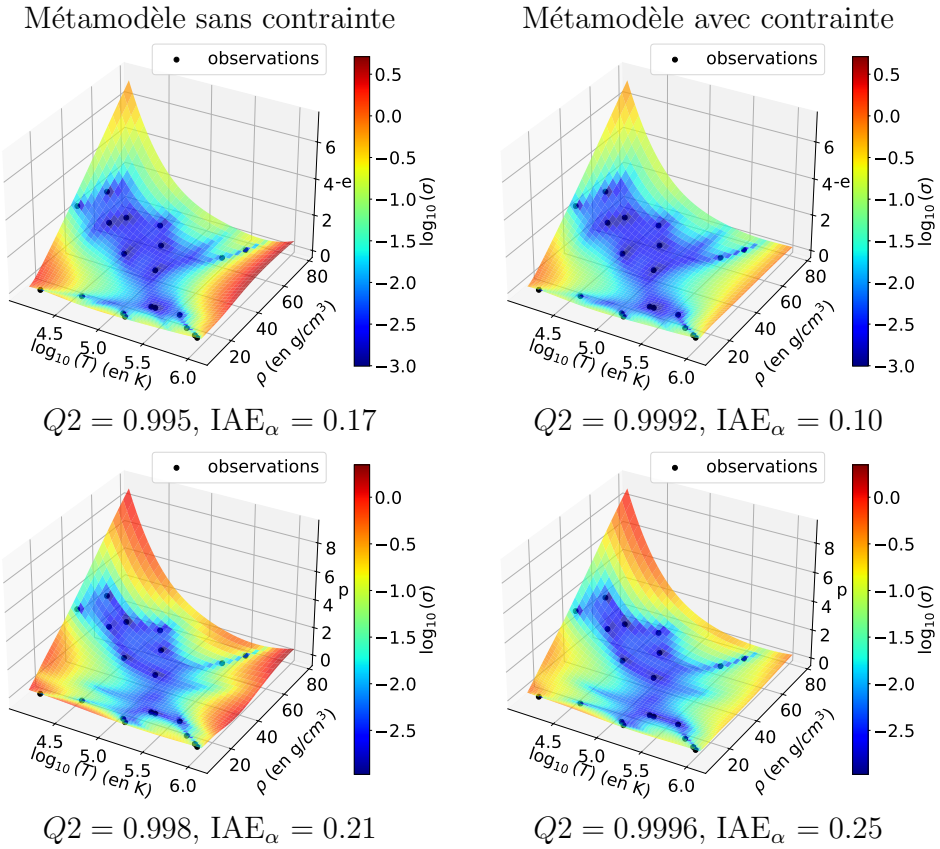


Figure 1: Moyenne et écart du type du processus gaussien approchant les variables adimensionnées ($-e$ et p) pour un modèle avec contrainte et un modèle sans contrainte avec 20 observations (points noirs). Les figures du haut, resp. du bas, representent e, resp. p.

Nous évaluons les résultats sur une grille régulière de points $(\tilde{x}_i, \tilde{y}_i)$. Pour évaluer la qualité de la prédiction, nous utilisons le $Q2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^n (\tilde{y}_i - \hat{y}(\tilde{x}_i))^2}{\frac{1}{n}\sum_{i=1}^n (\tilde{y}_i - \frac{1}{n}\sum_{i=1}^n \tilde{y}_i)^2}$ qui doit être au plus proche de 1. Pour évaluer la qualité des intervalles de confiance, nous utilisons l'$\text{IAE}_\alpha = \int_0^1 |\hat{\Delta}(\alpha) - \alpha| d\alpha$ avec $\hat{\Delta}(\alpha) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\tilde{y}_i \in \text{PI}_\alpha(\tilde{x}_i)}$ avec $\text{PI}_\alpha(\tilde{x}_i)$ l'intervalle de confiance de niveau α. La fonction $\hat{\Delta}(\alpha)$ doit être le plus proche possible de l'identité et donc l'IAE_α proche de 0.

Nous pouvons voir sur la figure 1 le résultat obtenu en utilisant comme données 20 observations choisies aléatoirement. Le niveau de surface représente la moyenne du processus et les couleurs l'écart type. La variance est plus élevée dans les zones avec peu d'observations dans le cas d'un modèle sans contrainte mais l'IAE_α est du même ordre de grandeur dans les deux cas. De plus, pour des températures élevées, l'allure de la courbe moyenne de e et de p diffère et l'indice $Q2$ est meilleur dans le cas avec contrainte.

4 Conclusion

Nous avons ici étudié une méthode permettant d'ajouter des contraintes en inégalité au modèle de la régression par processus gaussien en projetant un processus gaussien sur une famille de B-splines. Il est alors possible d'écrire sous la forme d'un nombre fini d'inégalités les conditions pour que le processus projeté soit borné, de dérivée bornée ou de dérivée seconde bornée. Cependant, dans certains cas, ces inégalités ne sont pas linéaires. Il est alors plus complexe d'utiliser l'algorithme d'échantillonnage permettant de calculer la loi du processus conditionnellement aux contraintes d'inégalités. Ainsi, nous avons proposé d'approximer ces inégalités par des inégalités linéaires. Enfin, nous avons appliqué notre modèle sur un jeu de donnée d'équations d'état pour prendre en compte des contraintes en inégalité.

Bibliographie

Da Veiga, S. and Marrel, A. (2012). Gaussian process modeling with inequality constraints. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 21, pages 529–555.

Eliezer, S., Ghatak, A. K., Hora, H., and Teller, E. (1986). *An introduction to equations of state: theory and applications*. Cambridge University Press Cambridge.

Kunoth, A., Lyche, T., Sangalli, G., Serra-Capizzano, S., Lyche, T., Manni, C., and Speleers, H. (2018). Foundations of spline theory: B-splines, spline approximation, and hierarchical refinement. *Splines and PDEs: From Approximation Theory to Numerical Linear Algebra: Cetraro, Italy 2017*, pages 1–76.

López-Lopera, A. F., Bachoc, F., Durrande, N., and Roustant, O. (2018). Finite-dimensional gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):1224–1255.

Maatouk, H. (2015). *Correspondance entre régression par processus gaussien et splines d'interpolation sous contraintes linéaires de type inégalité. Théorie et applications.* Theses, Ecole Nationale Supérieure des Mines de Saint-Etienne.

Maatouk, H., Rullière, D., and Bay, X. (2024). Bayesian analysis of constrained gaussian processes. *Bayesian Analysis*, 1(1):1–30.

Pakman, A. and Paninski, L. (2014). Exact hamiltonian monte carlo for truncated multivariate gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542.

Schmidt, J. W. and Heß, W. (1988). Positivity of cubic polynomials on intervals and positive spline interpolation. *BIT Numerical Mathematics*, 28:340–352.

Swiler, L. P., Gulian, M., Frankel, A. L., Safta, C., and Jakeman, J. D. (2020). A survey of constrained gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing*, 1(2).

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Block-Additive Gaussian Processes under Monotonicity Constraints

Mathis Deronzier[1,*], Andrés F. López-Lopera[2], François Bachoc[1,5], Olivier Roustant[3] and Jérémy Rohmer[4]

[1]*Institut de Mathématiques de Toulouse (IMT), Univ. Paul Sabatier, F-31062 Toulouse, France.*
[2]*Univ. Polytechnique Hauts-de-France, CERAMATHS, F-59313 Valenciennes, France.*
[3]*IMT, UMR5219 CNRS, INSA, F-31077 Toulouse cédex 4, France.*
[4]*BRGM, 3 avenue Claude Guillemin, F-45060 Orléans cédex 2, France.*
[5]*Institut Universitaire de France (IUF).*
[*]*Presenting author*

Résumé. Nous généralisons le cadre des processus gaussiens additifs contraints afin de gérer les interactions entre les variables d'entrée tout en imposant des contraintes de monotonie selon chacune des variables d'entrées. La structure bloc-additive du modèle est particulièrement adaptée en présence d'interactions, tout en conservant une complexité algorithmique maîtrisée. De plus, nous développons un algorithme séquentiel, MaxMod, pour la sélection du modèle (c'est-à-dire le choix des variables d'entrée actives et des blocs). Nous accélérons nos implémentations grâce à des calculs matriciels efficaces et à des expressions explicites des critères intervenant dans MaxMod. La performance et la scalabilité de notre méthodologie sont illustrées par plusieurs exemples numériques en dimensions allant jusqu'à 120, ainsi que par une application réelle en 5D sur les inondations côtières, où l'interprétabilité est renforcée par la sélection des blocs.

Mots-clés. Processus gaussiens, contraintes d'inégalité, modèle bloc-additif, allocation de nœuds, plans séquentiels, fonctions chapeau multidimensionnelles

Abstract. We generalize the additive constrained Gaussian process framework to handle interactions between input variables while enforcing monotonicity constraints everywhere on the input space. The block-additive structure of the model is particularly suitable in the presence of interactions, while maintaining tractable computations. In addition, we develop a sequential algorithm, MaxMod, for model selection (i.e., the choice of the active input variables and of the blocks). We speed up our implementations through efficient matrix computations and thanks to explicit expressions of criteria involved in MaxMod. The performance and scalability of our methodology are showcased with several numerical examples in dimensions up to 120, as well as in a 5D real-world coastal flooding application, where interpretability is enhanced by the selection of the blocks.

Keywords. Gaussian processes, inequality constraints, block-additive model, knot allocation, sequential designs, multi-dimensional hat functions

1 Introduction

Processus Gaussiens (PGs) contraints. Les PGs sont un outil fondamental des modèles bayésiens non paramétriques, avec de nombreux avantages théoriques et algorithmiques. L'intégration de contraintes (ex. bornes, monotonie, convexité) améliore la précision des prédictions et la qualité des incertitudes estimées.

Dans (Deronzier et al 2025), nous adaptons le cadre des PGs finis dimensionnels introduit dans (Maatouk et Bay 2017) pour traiter la contrainte de monotonie. En particulier, nous travaillons sur la projection de nos priors gaussiens sur des espaces de fonctions finis dimensionnels engendrés par des fonctions de base "chapeau" multidimensionnelles, supportées à l'intérieur d'un parallépipède décrit par les nœuds d'une grille. Le nombre de fonctions dans la base dépend directement de la taille de la grille multidimensionnelle et donc croit exponentiellement avec la dimension de l'espace. Ce modèle est donc sujet au "fléau de la dimension", une méthode séquentielle appelée MaxMod permettant de choisir les localisation des nœuds et les dimensions actives est développée dans (Bachoc et al 2022).

PGs additifs contraints. Dans (López-Lopera et al 2022), une hypothèse d'additivité est faite sur la fonction cible:

$$y(x_1, \dots, x_D) = y_1(x_1) + \cdots + y_D(x_D). \tag{1}$$

MaxMod y est adapté à la sélection de variables et l'allocation des nœuds. Ce cadre permet de significativement augmenter la dimension des fonctions traitées ($D \sim 100$). Cependant cette hypothèse ne permet pas de gérer les fonctions ayant des interactions entre les variables.

Extension aux PGs bloc-additifs (PGsba). Nous proposons dans (Deronzier et al 2025) une approche généralisant (Bachoc et al 2022) et (López-Lopera et al 2022), permettant d'introduire des interactions via une décomposition bloc-additive:

$$y(x_1, \dots, x_D) = y_1(\boldsymbol{x}_{\mathcal{B}_1}) + \cdots + y_B(\boldsymbol{x}_{\mathcal{B}_B}). \tag{2}$$

Les blocs $\mathcal{P} := \{\mathcal{B}_1, \dots, \mathcal{B}_B\}$ forment une sous-partition des variables. Ce modèle permet de capturer des interactions tout en conservant une complexité raisonnable.

L'inférence de la structure bloc-additive est un défi, et à notre connaissance, notre approche est la première à traiter ce problème dans un cadre de PGs contraints.

Résumé des contributions. Nous apportons plusieurs contributions :

1) Introduction d'un cadre théorique pour les PGsba contraints, avec des bases adaptées et des résultats analytiques sur les matrices de changement de base.

2) Extension de MaxMod aux PGsba, avec une expression explicite du critère d'optimisation.

3) Décomposition en prédicteurs par bloc y_i permettant une meilleure interprétabilité. Une application aux inondations côtières illustre l'intérêt physique de cette approche (Deronzier et al 2025)[section 5.4 et appendice E].

4) Validation numérique jusqu'à $D = 120$, montrant l'efficacité de MaxMod pour identifier les variables influentes et optimiser les modèles (Deronzier et al 2025)[sections 5.2 et 5.3].

5) Mise à disposition d'un code dans la bibliothèque R `lineqGPR` [5].

2 Les PGsba finis dimensionnels

On suppose dans cette section la sous-partition $\mathcal{P} = \{\mathcal{B}_1, \cdots, \mathcal{B}_B\}$ fixée.

Définition d'un PG bloc-additif (PGba). Pour chaque bloc $\mathcal{B}_j$, $1 \leq j \leq B$, nous considérons un PG centré $\{Y_j(\boldsymbol{x}),\ \boldsymbol{x} \in [0,1]^{|\mathcal{B}_j|}\}$ avec noyau k_j. Le PGba est alors défini comme la somme des PGs associés aux blocs:

$$Y^{\mathcal{P}}(\boldsymbol{x}) = Y_1(\boldsymbol{x}_{\mathcal{B}_1}) + \cdots + Y_B(\boldsymbol{x}_{\mathcal{B}_B}). \tag{3}$$

On suppose les $(Y_j)_{1\leq j\leq B}$ indépendants eux, alors $Y^{\mathcal{P}}$ est également un GP centré dont on peut exprimer le noyau $k_{\mathcal{P}}$.

Construction des fonctions chapeau. Dans le cas de la dimension un, une base de fonctions chapeau $\{\phi_1, \cdots, \phi_m\}$ peut être définie à partir d'une subdivision $s := (t_1, \cdots, t_m)$ de $[0,1]$: $0 = t_1 < \cdots < t_m = 1$ (voir la figure 1). On peut encore définir la projection d'une fonction $y : [0,1] \to \mathbb{R}$ sur l'espace engendré par la base de fonctions comme la fonction affine sur les intervalles $[t_i, t_{i+1}]$ interpollant y aux points $t_1, \cdots, t_m$.

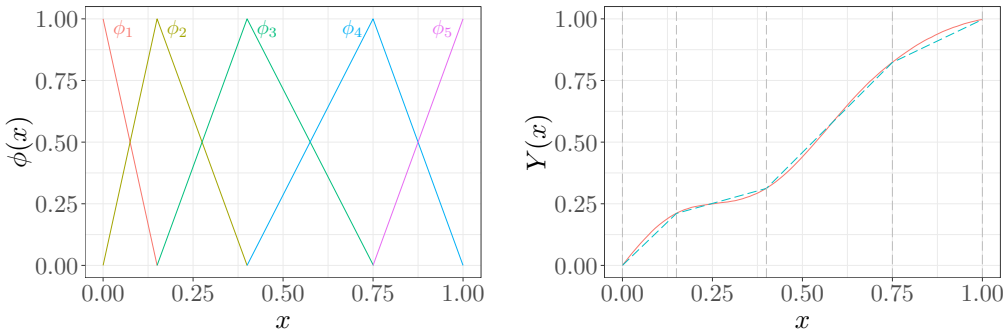


Figure 1: La première figure représente une base de fonctions chapeau générée à partir de la subdivision $s = (0, 0.2, 0.4, 0.75, 1)$, et la seconde un exemple de la projection (en bleu) d'une fonction monotone (en rouge) sur l'espace vectoriel engendré par la base.

Les familles multidimensionnelles sont obtenues en tensorisant des bases unidimensionnelles (voir figure 2). De façon générale, étant données des subdivisions $\mathcal{S} := (s^{(1)}, \cdots, s^{(D)})$ de $[0,1]$, $s^{(i)} := \{t^{s^{(i)}}, \cdots, t^{s^{(i)}}_{m^{(i)}}\}$. Elles définissent des bases $\beta_{s^{(1)}}, \cdots, \beta_{s^{(D)}}$, $\beta_{s^{(i)}} := \{\phi_1^{s^{(i)}}, \cdots, \phi^{s^{(i)}}_{m^{(i)}}\}$. Pour chaque block $\mathcal{B}_j$, on peut définir un ensemble d'indices:

$$\mathcal{L}^{\mathcal{S}}_{\mathcal{B}_j} = \prod_{i\in\mathcal{B}_j} \left\{1, \ldots, m^{(i)}\right\} = \left\{\underline{\ell}_j = (\underline{\ell}_{j,i})_{i\in\mathcal{B}_j},\ 1 \leq \underline{\ell}_{j,i} \leq m^{(i)},\ \forall i \in \mathcal{B}_j\right\}. \tag{4}$$

Et à chaque indice $\underline{\ell}_j$ dans $\mathcal{L}^{\mathcal{S}}_{\mathcal{B}_j}$, la coordonnée d'un point de la grille multidimensionnelle générée par les subdivisions $(s^{(i)})_{i\in\mathcal{B}_j}$, correspond une fonction chapeau (voir figure 2), $\phi_{\underline{\ell}_j} : [0,1]^D \to \mathbb{R}$

$$\phi_{\underline{\ell}_j}(\boldsymbol{x}) = \prod_{i\in\mathcal{B}_j} \phi^{s^{(i)}}_{\underline{\ell}_{j,i}}(x_i). \tag{5}$$

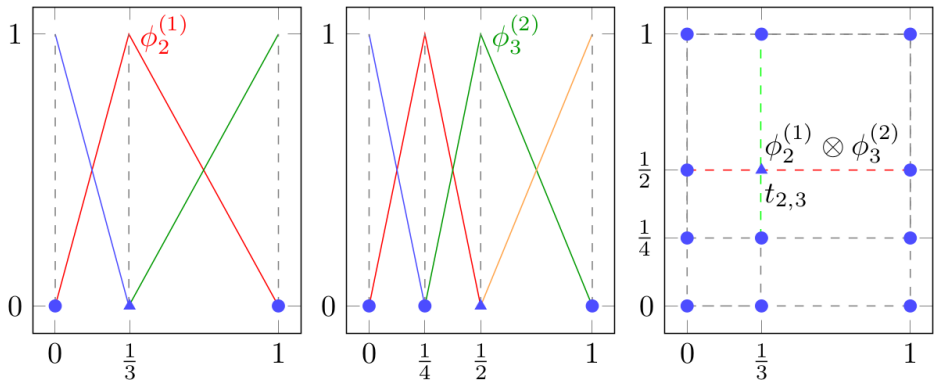


Figure 2: Les deux premières figures représentent des bases obtenues à partir des subdivisions $\mathcal{S} = ((0, 1/3, 1), (0, 1/4, 1/2, 1))$, et la troisième une bases 2-dimensionnelle obtenue par tensorisation des deux premières, la fonction $\phi_{(2,3)}$ est affichée sur cette dernière.

On peut définir une projections $P_{\mathcal{P}}^{\mathcal{S}}$ sur l'espace $E_{\mathcal{P}}^{\mathcal{S}}$ engendré par les fonctions $(\phi_{\underline{\ell}_j})_{\underline{\ell}_j \in \mathcal{L}_{\mathcal{B}_j}^{\mathcal{S}}, 1 \leq j \leq B}$. Pour une fonction $f = \sum_j f_j$ bloc-additive selon la structure induite par $\mathcal{P} = \{\mathcal{B}_1, \cdots, \mathcal{B}_B\}$:

$$P_{\mathcal{P}}^{\mathcal{S}}\left(\sum_{j=1}^{B} f_j\right) = \sum_{j=1}^{B} \sum_{\underline{\ell}_j \in \mathcal{L}_{\mathcal{B}_j}^{\mathcal{S}}} f_j(t_{\underline{\ell}_j}) \phi_{\underline{\ell}_j}. \tag{6}$$

Et donc définir le processus gaussien fini dimensionnel $\widetilde{Y}_{\mathcal{P}}^{\mathcal{S}}$:

$$\widetilde{Y}_{\mathcal{P}}^{\mathcal{S}}(\boldsymbol{x}) = P_{\mathcal{P}}^{\mathcal{S}}(Y^{\mathcal{P}})(\boldsymbol{x}) = \sum_{j=1}^{B} \sum_{\underline{\ell}_j \in \mathcal{L}_{\mathcal{B}_j}^{\mathcal{S}}} Y_j(t_{\underline{\ell}_j}) \phi_{\underline{\ell}_j}(\boldsymbol{x}). \tag{7}$$

3 Conditionnement de $\widetilde{Y}_{\mathcal{P}}^{\mathcal{S}}$

On cherche maintenant à conditionner notre PGba fini dimensionnel $\widetilde{Y}_{\mathcal{P}}^{\mathcal{S}}$ à la condition d'interpolation de la fonction cible: $\widetilde{Y}_{\mathcal{P}}^{\mathcal{S}}(\boldsymbol{X}) = \boldsymbol{Y} + \boldsymbol{\epsilon}$. Ici, $\boldsymbol{X}$ et $\boldsymbol{Y}$ sont respectivement les vecteur des points d'observation et les valeurs observées à ces points, de sorte que $y(\boldsymbol{X}_i) = \boldsymbol{Y}_i$ pour $i = 1, \cdots, n$. Le terme $\boldsymbol{\epsilon}$ est un bruit gaussien de loi $\mathcal{N}(0, \tau \boldsymbol{I}_n)$. La partie difficile est de conditionner à la contrainte fonctionnelle de monotonie qu'on écrira $\widetilde{Y}_{\mathcal{P}}^{\mathcal{S}} \in \mathcal{C}$. $\mathcal{C}$ est ici l'ensemble des fonctions monotones. D'après l'équation (7), on peut réécrire

$$\widetilde{Y}_{\mathcal{P}}^{\mathcal{S}}(\boldsymbol{x}) = \boldsymbol{\Phi}_{\mathcal{P}}^{\mathcal{S}}(\boldsymbol{x})^{\top} \boldsymbol{\xi}.$$

Où $\boldsymbol{\Phi}_{\mathcal{P}}^{\mathcal{S}} := (\phi_{\underline{\ell}_j})_{\underline{\ell}_j \in \mathcal{L}_{\mathcal{B}_j}^{\mathcal{S}}, 1 \leq j \leq B}$ est le vecteur des fonctions chapeau et $\boldsymbol{\xi}$ un vecteur gaussien dont on connait la loi puisqu'il est composé des réalisations des priors gaussiens Y_j introduits au début de la section 2. Comme $\boldsymbol{\Phi}_{\mathcal{P}}^{\mathcal{S}}$ est déterministe, on peut réécrire:

$$\left(\widetilde{Y}_{\mathcal{P}}^{\mathcal{S}}(\boldsymbol{x}) \middle| \widetilde{Y}_{\mathcal{P}}^{\mathcal{S}}(\boldsymbol{X}) + \boldsymbol{\epsilon} = \boldsymbol{Y}, \widetilde{Y}_{\mathcal{P}}^{\mathcal{S}} \in \mathcal{C}\right) = \boldsymbol{\Phi}_{\mathcal{P}}^{\mathcal{S}}(\boldsymbol{x})^{\top} \left(\boldsymbol{\xi} \middle| \widetilde{Y}_{\mathcal{P}}^{\mathcal{S}}(\boldsymbol{X}) = \boldsymbol{Y} + \boldsymbol{\epsilon}, \widetilde{Y}_{\mathcal{P}}^{\mathcal{S}} \in \mathcal{C}\right).$$

L'introduction de l'espace fini dimensionnel $E_{\mathcal{P}}^{\mathcal{S}}$, de dimension $|\mathcal{L}_{\mathcal{P}}^{\mathcal{S}}|$, intervient ici puisque que nous avons une équivalence entre l'appartenance du processus gaussien $\widetilde{Y}_{\mathcal{P}}^{\mathcal{S}}$ aux fonctions

monotones et celle du vecteur gaussian $\boldsymbol{\xi}$ à un sous espace de $\mathbb{R}^{|\mathcal{L}_{\mathcal{P}}^{\mathcal{S}}|}$ caractérisé par des contraintes linéaires:

$$\widetilde{Y}_{\mathcal{P}}^{\mathcal{S}} \in \mathcal{C} \iff \left(\boldsymbol{\Phi}_{\mathcal{P}}^{\mathcal{S}}\right)^{\top} \boldsymbol{\xi} \in \mathcal{C} \iff \boldsymbol{\xi} \in \mathcal{C}' \iff \boldsymbol{\Lambda}\boldsymbol{\xi} \leq 0. \tag{8}$$

Dans le cas de la dimension 1, on peut facilement exprimer la matrice $\boldsymbol{\Lambda}$, en effet une fonction affine par morceau est croissante si et seulement si la suite de ses valeurs aux nœuds est croissante (voir figure 1). Autrement dit, pour une fonction $f = \sum_{i=1}^{m} a_i \phi_i$ définie à partir de fonctions chapeau comme en figure 1, on a l'équivalence suivante:

$$f \text{ est croissante} \iff a_i \leq a_{i+1},\ \forall i \in \{1, \ldots, m-1\}. \tag{9}$$

Finalement, on construit notre prédicteur à partir du mod (max of distribution) sur $\mathcal{C}'$ du vecteur gaussien conditionné $(\boldsymbol{\xi} | \widetilde{Y}_{\mathcal{P}}^{\mathcal{S}}(\boldsymbol{X}) = \boldsymbol{Y} + \boldsymbol{\epsilon}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. C'est la solution du problème d'optimisation quadratique:

$$\widehat{\boldsymbol{\xi}} = \underset{\boldsymbol{x} \in \mathbb{R}^{|\mathcal{L}_{\mathcal{P}}^{\mathcal{S}}|},\, \boldsymbol{\Lambda}\boldsymbol{x} \leq 0}{\arg\min} \ (\boldsymbol{x} - \boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \tag{10}$$

Le prédicteur sera finalement donné par l'équation:

$$\widehat{Y}_{\mathcal{P}}^{\mathcal{S}} = (\boldsymbol{\Phi}_{\mathcal{P}}^{\mathcal{S}})^{\top} \widehat{\boldsymbol{\xi}}. \tag{11}$$

4 Construction séquentielle des PGsba contraints via MaxMod

Dans la section précédente, nous avons construit le prédicteur $\widehat{Y}_{\mathcal{P}}^{S}$ (11). Cette construction dépend des subdivisions $\mathcal{S}$, de la sous-partition $\mathcal{P}$ et de l'ensemble convexe $\mathcal{C}'$. Le coût de calcul du prédicteur, solution de (10), augmente avec le nombre total de fonctions de base $|\mathcal{L}_{\mathcal{P}}^{\mathcal{S}}|$. Cette section introduit une méthode itérative pour sélectionner $\mathcal{P}$ et $\mathcal{S}$.

Choix possibles pour mettre à jour la sous-partition et les subdivisions. L'approche consiste à mettre à jour séquentiellement $\mathcal{P}$ et $\mathcal{S}$. À l'initialisation, $\mathcal{P} = \{\}$ et $\mathcal{S} = ((), \cdots, ())$. Trois options sont considérées pour mettre à jour $\mathcal{P}$ et $\mathcal{S}$:

- **ACTIVATION :** Activer une variable i en définissant $s^{(i)} := (0, 1)$ et intégrer le bloc $\{i\}$ à $\mathcal{P}$.
- **RAFFINEMENT :** Ajouter un nœud $t \in [0, 1]$ à une subdivision existante $s^{(i)}$.
- **FUSION :** Fusionner deux blocs $\mathcal{B}_a$ et $\mathcal{B}_b$ en un seul.

Ces options définissent l'ensemble des mises à jour possibles $\mathcal{M}^{\star}(\mathcal{S}, \mathcal{P})$. Le choix de la meilleure option sera déterminé à partir de la valeur d'un critère.

Construction du critère MaxMod Le critère MaxMod combine deux sous-critères :

- **Modification L^2 (L2Mod) :** Mesure la différence entre deux prédicteurs successifs :

$$\text{L2Mod}((\mathcal{S},\mathcal{P}),(\mathcal{S}^\star,\mathcal{P}^\star)) = \int_{[0,1]^D} \left(\widehat{Y}^{\mathcal{S}^\star}_{\mathcal{P}^\star}(x) - \widehat{Y}^{\mathcal{S}}_{\mathcal{P}}(x)\right)^2 dx. \tag{12}$$

  Une expression algébrique de $\text{L2Mod}((\mathcal{S},\mathcal{P}),(\mathcal{S}^\star,\mathcal{P}^\star))$ est donnée dans l'article [1].

- **Erreur quadratique (SE) :** Évalue la précision du prédicteur :

$$\text{SE}(\mathcal{S}^\star,\mathcal{P}^\star) = \left\|\widehat{Y}^{\mathcal{S}^\star}_{\mathcal{P}^\star}(\boldsymbol{X}) - \boldsymbol{Y}\right\|^2. \tag{13}$$

Le critère final est défini par :

$$\mathcal{K}(\mathcal{S}^\star,\mathcal{P}^\star) = \frac{\text{L2Mod}((\mathcal{S},\mathcal{P}),(\mathcal{S}^\star,\mathcal{P}^\star))}{(|\mathcal{L}^{\mathcal{S}^\star}_{\mathcal{P}^\star}| - |\mathcal{L}^{\mathcal{S}}_{\mathcal{P}}|)^\alpha \text{SE}(\mathcal{S}^\star,\mathcal{P}^\star)^\gamma}. \tag{14}$$

où α et γ sont des paramètres de pondération. Des tests ont montré que les valeurs optimales sont $\alpha = 1.4$ et $\gamma = 0.5$.

Algorithme MaxMod. L'algorithme Maxmod est présenté ci-dessous:

Algorithm 1 MaxMod

Require: Observations $(\boldsymbol{X},\boldsymbol{Y})$, critères d'arrêt $\epsilon_1,\epsilon_2 \in (0,1)$, nombre max d'itérations M
Ensure: Subdivision $\mathcal{S}$, partition $\mathcal{P}$ et prédicteur $\widehat{Y}^S_{\mathcal{P}}$

1: Initialisation: $\mathcal{S} = ((),\ldots,())$, $\mathcal{P} = \{\}$, $c_1 = 2\epsilon_1$, $c_2 = 2\epsilon_2$, $i = 0$
2: **while** $c_1 > \epsilon_1$ et $c_2 > \epsilon_2$ et $i \leq M$ **do**
3: Sélection de $(\mathcal{S}^\star,\mathcal{P}^\star)$ maximisant $\mathcal{K}$
4: $c_1 := \text{L2Mod}((\mathcal{S},\mathcal{P}),(\mathcal{S}^\star,\mathcal{P}^\star))$, $c_2 := \text{SE}(\mathcal{S}^\star,\mathcal{P}^\star)$
5: $\mathcal{S} = \mathcal{S}^\star$, $\mathcal{P} = \mathcal{P}^\star$, $i = i+1$
6: **end while**
7: Calcul du prédicteur $\widehat{Y}^{\mathcal{S}}_{\mathcal{P}}$ selon (11)
8: **return** $(\mathcal{S},\mathcal{P},\widehat{Y}^{\mathcal{S}}_{\mathcal{P}})$

Bibliographie

[1] Deronzier, M. López-Lopera, AF. Bachoc, F. et Roustant,O. (2025) *Block-Additive Gaussian Processes under Monotonicity Constraints*, arXiv preprint, https://arxiv.org/abs/2407.13402.

[2] Bachoc, F. López-Lopera, AF. et Roustant, O. (2022) *Sequential construction anddimension reduction of gaussian processes under inequality constraints*, SIAM.

[3] López-Lopera, AF. Bachoc, F. et Roustant, O. (2022) *High-dimensional additive Gaussianprocesses under monotonicity constraints*, Advances in Neural Information Processing Systems, **35**, 8041–8053.

[4] Maatouk, H. Bay, X. (2017) *Gaussian process emulators for computer experiments with inequality constraints*, Mathematical Geosciences **49**(5), 557–582.

[5] López-Lopera, A.F., Deronzier, M. (2022) *lineqGPR: Gaussian Process Regression with Linear Inequality Constraints*, R, package version 0.3.0. https://github.com/anfelopera/lineqGPR.

Résolution d'un problème inverse sur des expériences laser avec modèle génératif

Cécile Haberstich [1] & Claire Cannamela [1]

[1] *CEA, DAM, DIF, F-91297, Arpajon, France, {cecile.haberstich,claire.cannamela}@cea.fr*

Résumé. Ce travail vise à déterminer la densité des matériaux d'un objet en cours d'implosion à partir d'un petit nombre de radiographies. Il s'agit d'un problème inverse inhabituel de reconstruction tomographique qui est "mal posé" et nécessite donc une information a priori sur la solution pour maîtriser les instabilités numériques. Dans de nombreux travaux, l'utilisation de modèles génératifs pour intégrer cette information a priori par régularisation a prouvé son efficacité. En particulier, les Generative Adversarial Networks (GAN) permettent, via l'apprentissage d'un réseau de neurones, de relier un espace latent gaussien avec l'espace des images d'intérêt. La résolution du problème inverse se fait alors dans l'espace latent pénalisé par la norme au carré des vecteurs afin de conserver les propriétés gaussiennes. Ici, nous proposons une nouvelle manière de régulariser cet espace latent basé sur la distance de Wasserstein.

Mots-clés. Problème inverse, modèle génératif, espace latent, régularisation Wasserstein

Abstract. This work aims to determine the density of materials in an object undergoing implosion from a small number of radiographs. This is an unusual inverse problem of tomographic reconstruction which is ill-posed and therefore requires prior information about the solution to control numerical instabilities. The use of generative models to incorporate this prior information through regularization has proven effective in many works. In particular, Generative Adversarial Networks (GAN) allow for a Gaussian latent space to be connected with the space of images of interest through the learning of a neural network. The inverse problem is then solved in the penalized latent space, with the square of vector norms to preserve Gaussian properties. In this work, we propose a new method for regularizing this latent space based on the Wasserstein distance.

Keywords. Inverse problem, generative model, latent space, Wasserstein regularization

1 Introduction

Ce travail s'inscrit dans le cadre de la résolution d'un problème inverse de type *tomographie*. Le lecteur pourra se référer à l'ouvrage de Arridge (2019), référence complète sur le sujet. C'est une technique permettant de reconstruire une image 2D ou 3D d'un objet à partir d'un ensemble de projections (ou de mesures) effectuées à l'extérieur de l'objet. Dans ce problème inverse, les projections sont des informations partielles de l'image à reconstruire, et on doit résoudre l'inverse de l'opération de projection pour obtenir l'image complète. Par

exemple, dans l'étude d'implosion présentée à la section 5, il s'agit de déterminer les cartes de température et de densité de matériaux qui décrivent l'état de la matière à partir d'images obtenues en utilisant des rayons X.

De manière plus formelle, l'objectif est de retrouver une image $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ à partir de mesures $\mathbf{y} \in \mathcal{Y} \subset \mathbb{R}^n$ éventuellement bruitées, de bruit ε :

$$\mathbf{y} = \mathcal{A}(\mathbf{x}) + \varepsilon, \tag{1}$$

où $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ est un opérateur inconnu. La solution de ce problème inverse est donnée par :

$$\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathcal{A}(\mathbf{x}) - \mathbf{y}\|^2. \tag{2}$$

Les problèmes inverses, notamment ceux du type *tomographie* où $d \gg n$, sont souvent mal posés, *i.e.* l'une au moins des conditions d'existence, d'unicité ou de continuité des solutions n'est pas vérifiée, et/ou mal conditionnés, *i.e.* les solutions sont très sensibles à l'incertitude de mesure.

2 Approches classiques et formalisme bayésien

Lorsque le problème inverse (1) est mal posé, il est courant d'ajouter un terme de régularisation qui pénalise les solutions trop irrégulières :

$$\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathcal{A}(\mathbf{x}) - \mathbf{y}\|^2 + \lambda R(\mathbf{x}), \tag{3}$$

avec λ un paramètre de régularisation qui permet un compromis entre l'attache aux données, avec le terme $\|\mathcal{A}(\mathbf{x}) - \mathbf{y}\|^2$, et la pénalisation $R(\mathbf{x})$. Le choix de $\lambda > 0$ est souvent effectué par validation croisée et permet de résoudre le problème de mauvais conditionnement de $\mathcal{A}$.

Une première catégorie de régularisations comporte les fonctions différentiables et convexes telles que la régularisation de Tikhonov (1963), qui favorise les solutions lisses en encourageant la reconstruction à être petite pour la norme 2. Elle s'exprime selon $R(\mathbf{x}) = \|\nabla\mathbf{x}\|_2^2$. Une deuxième catégorie de régularisations a vu le jour avec les pénalisations non lisses, comme la variation totale (notée TV), introduite par Rudin et al. (1963) qui permet de conserver les discontinuités. La régularisation par TV peut s'écrire sous la forme $R(\mathbf{x}) = \|\nabla\mathbf{x}\|_1^2$.

Pour les problèmes inverses mal posés, la formulation probabiliste apporte plus de flexibilité sur le type d'images à reconstruire en considérant une famille de solutions compatibles avec les données observées. L'image $\mathbf{x}$ est alors supposée être une réalisation d'un vecteur aléatoire $\mathbf{X}$ dont la densité de probabilité est $p(\mathbf{x})$. Le vecteur aléatoire $\mathbf{Y}$ est relié à $\mathbf{X}$ par le modèle direct $\mathbf{Y} = \mathcal{A}(\mathbf{X}) + \varepsilon$.

Dans le cadre bayésien, la densité $p(\mathbf{x})$ est appelée densité *a priori* et reflète l'état des connaissances que l'on a sur l'image à reconstruire. Lorsque cette densité est combinée aux données observées via la fonction de vraisemblance $p(\mathbf{y}|\mathbf{x})$ (connue), la densité de probabilité *a posteriori* de $\mathbf{X}$ est donnée par : $p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$.

Néanmoins, le choix d'une distribution *a priori* $p(\mathbf{x})$ pour représenter des images est délicat, car les images sont souvent complexes et variées. Une approche actuelle propose de construire $p(\mathbf{x})$ à l'aide d'un modèle génératif basé sur des réseaux de neurones. Cela fera l'objet de la section 3.

3 Modèles génératifs et WGAN-GP

Récemment, les modèles génératifs basés sur des réseaux de neurones ont démontré leur capacité à approcher une distribution complexe d'images. Parmi les modèles génératifs les plus populaires, on peut citer les auto-encodeurs variationnels (VAE), introduits par Kingma et Welling en 2014, les modèles de diffusion proposés par Feizi et al. en 2018, et les GAN (Generative Adversarial Networks), initiés par Goodfellow et al. en 2014.

Un modèle génératif, paramétré par θ et noté G_θ, cherche à apprendre $p(\mathbf{x})$, la distribution empirique de la base de données d'images notée $\{\mathbf{x}^i\}_{i=1}^N$, au travers de la distribution μ_θ, distribution de probabilité que le modèle est capable de produire. Pour ce faire, on considère qu'une donnée observée $\mathbf{x}^i \in \mathcal{X}$ peut être vue comme le résultat d'une transformation d'un vecteur aléatoire, appelé vecteur latent et noté $\mathbf{z}^i \in \mathbb{R}^k$ avec $k \ll d$. Cette transformation est effectuée via le modèle génératif $G_\theta : \mathbb{R}^k \to \mathcal{X}$ tel que $\mathbf{x}^i = G_\theta(\mathbf{z}^i),\ \mathbf{z}^i \sim p_{\mathbf{z}}(\mathbf{z})$ avec $p_{\mathbf{z}}(\mathbf{z})$ une loi dont il est simple de générer des réalisations, par exemple une gaussienne $\mathcal{N}(0, I_k)$. L'apprentissage du modèle génératif consiste alors à trouver les paramètres θ permettant de produire des réalisations aussi proches que possible de la distribution d'intérêt. Cela se fait en cherchant le vecteur θ qui minimise une fonction de perte $\mathcal{L}$, par exemple la divergence de Kullback-Leibler, mesurant l'écart entre la distribution empirique $p(\mathbf{x})$ des données et la distribution de probabilité μ_θ produite par le modèle $G_\theta(\mathbf{z})$: $\inf_{\theta\in\Theta} \mathcal{L}(\mu_\theta, p(\mathbf{x}))$.

Dans ce travail, nous nous sommes intéressés aux GAN. Ces derniers sont composés de deux réseaux de neurones. Le premier réseau est le générateur G_θ qui génère des images réalistes à partir d'un bruit aléatoire gaussien. Le second réseau est le discriminateur D_γ qui doit différencier les vraies images des fausses. Les réseaux sont entrainés simultanément et de manière compétitive (ou adverse), *i.e.* l'un essaie de maximiser son gain tandis que l'autre essaie de minimiser sa perte. En effet, le générateur G_θ essaie de tromper le discriminateur en produisant des exemples de plus en plus réalistes. Le terme $D_\gamma(G_\theta(\mathbf{z}))$ doit donc être proche de 1 pour tout $\mathbf{z} \in \mathbb{R}^k$. Tandis que le discriminateur $D_\gamma : \mathcal{X} \to [0, 1]$ cherche à différencier les images de la base de données $\{\mathbf{x}^i\}_{i=1}^N$ de celles générées par $G_\theta(\mathbf{z})$. Le terme $D_\gamma(G_\theta(\mathbf{z}))$ doit ainsi être proche de 0 pour tout $\mathbf{z}$, et $D_\gamma(\mathbf{x}^i)$ doit être proche de 1 pour $i = 1, \ldots, N$. La fonction de perte globale est alors donnée par :

$$\min_{\theta\in\Theta} \max_{\gamma\in\Gamma} \mathbb{E}_{\boldsymbol{X}\sim p(\boldsymbol{x})}[\log D_\gamma(\boldsymbol{X})] + \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,I_k)}[\log(1 - D_\gamma(G_\theta(\mathbf{z}))], \tag{4}$$

où $\mathbf{X}$ suit la distribution $p(\mathbf{x})$ et $\mathbf{Z}$ suit la distribution $\mathcal{N}(0, I_k)$.

Les GAN ont cependant rapidement montré des limites. L'entraînement des GAN est plutôt instable. On peut citer par exemple le mode "collapse" dans lequel le générateur produit toujours la même image (un seul mode) ou encore le mode "dropping" dans lequel certains types d'images ne sont jamais générées (modes manquants).

Afin d'améliorer la stabilité de l'entraînement et de réduire les problèmes courants des GAN, Arjovsky et al. (2017) ont proposé une approche inspirée du transport optimal sous le nom de Wasserstein GAN (WGAN). Dans les WGAN, la fonction de perte pour l'apprentissage du générateur G_θ est la distance de Wasserstein W_1 définie par :

$$\mathrm{W}_1(\mu_\theta, p(\mathbf{x})) = \inf_{\gamma \in \Pi(\mu_\theta, p(\mathbf{x}))} \mathbb{E}_{(\mathbf{X}', \mathbf{X}) \sim \gamma} \left[\|\mathbf{X}' - \mathbf{X}\|_1 \right],$$

où $\mathbf{X}'$ et $\mathbf{X}$ sont des variables aléatoires de loi respectives μ_θ et $p(\mathbf{x})$, $\Pi(\mu_\theta, p(\mathbf{x}))$ est l'ensemble des distributions de probabilité conjointes $\gamma(\mathbf{X}', \mathbf{X})$ ayant μ_θ et $p(\mathbf{x})$ comme marginales.

Dans Santambrogio (2015), il est montré que le discriminateur optimal du WGAN satisfait $\|\nabla_{\mathbf{x}} D_\gamma(\mathbf{x})\| = 1$ (Lipschitz-1). Une version améliorée du WGAN, appelée le WGAN-GP (avec pénalité de gradient) a été proposée par Gulrajani er al. (2017) pour s'assurer que la contrainte de Lipschitz-1 est respectée. C'est cette dernière approche que nous avons utilisée dans ce travail lors de la résolution du problème inverse, qui fera l'objet de la section 4.

4 Résolution du problème inverse

Dans l'espace latent, le problème inverse est mieux posé, c'est pourquoi de nombreux travaux, tels que ceux de Ongie et al. (2020), passent par une formulation directe :

$$\hat{\mathbf{x}} = G_\theta(\hat{\mathbf{z}}) \quad | \quad \hat{\mathbf{z}} \in \arg\min_{\mathbf{z} \in \mathbb{R}^k} \|\mathcal{A}(G_\theta(\mathbf{z})) - \mathbf{y}\|^2. \tag{5}$$

Remarquons que cela revient à restreindre l'espace des solutions images à un espace engendré par le modèle génératif $\Omega = \{\hat{\mathbf{x}} = G_\theta(\mathbf{z}), \mathbf{z} \in \mathbb{R}^k\} \subset \mathbb{R}^d$ et que le problème inverse peut aussi s'écrire : $\hat{\mathbf{x}} = \arg\min_{\mathbf{x} \in \Omega} \|\mathcal{A}(\mathbf{x}) - \mathbf{y}\|^2$.

Cette résolution directe dans l'espace latent n'est pas suffisante pour nos applications de tomographie. L'espace des possibles pour trouver le vecteur latent optimal $\hat{\mathbf{z}}$ dans (5) est infini et la solution du problème inverse pourrait trop "s'écarter" d'une réalisation d'une gaussienne, entraînant la génération d'une image non réaliste par le modèle génératif. On cherche ainsi à contraindre la solution du problème inverse à rester proche d'une réalisation d'une gaussienne $\mathcal{N}(0, I_k)$ en ajoutant un terme de régularisation.

Une première approche possible est le "clipping", utilisée dans Paris et al. (2017), dans laquelle on cherche à maintenir $\mathbf{z}$ dans une certaine plage, le problème inverse devient :

$$\hat{\mathbf{x}}^{\text{clip}} = G_\theta(\hat{\mathbf{z}}^{\text{clip}}) \quad | \quad \hat{\mathbf{z}}^{\text{clip}} \in \arg\min_{\mathbf{z} \in \mathbb{R}^k \cap [-c,c]^k} \|\mathcal{A}(G_\theta(\mathbf{z})) - \mathbf{y}\|^2, \tag{6}$$

avec le paramètre $c \approx 3$ qui limite l'amplitude des valeurs latentes pendant l'optimisation en évitant que certaines dimensions deviennent trop grandes. Cette approche est très simple à implémenter mais pourrait empêcher une exploration suffisante de l'espace latent.

Une autre approche possible, utilisée dans Dumoulin et al (2018), consiste à imposer une pénalité de type ridge qui contraint la norme de $\mathbf{z}$ à ne pas devenir trop grande :

$$\hat{\mathbf{x}}^{\text{ridge}} = G_\theta(\hat{\mathbf{z}}^{\text{ridge}}) \quad | \quad \hat{\mathbf{z}}^{\text{ridge}} \in \arg\min_{\mathbf{z} \in \mathbb{R}^k} \|\mathcal{A}(G_\theta(\mathbf{z})) - \mathbf{y}\|^2 + \lambda_{\text{ridge}} \|\mathbf{z}\|_2^2, \tag{7}$$

où λ_{ridge} est un paramètre de régularisation. Ce type de régularisation favorise les composantes $\mathbf{z}$ à rester proche de 0. Cette approche est simple à mettre en œuvre, mais ne garantit pas que la distribution de $\mathbf{z}$ suive une gaussienne standard.

Des travaux tels que ceux de Kingma et Welling (2014) avec les auto-encodeurs variationnels, ont montré l'importance de régulariser l'espace latent à l'aide de la divergence de Kullback-Leibler. Une approche complémentaire, introduite par Tolstikhin et al. (2018) avec les auto-encodeurs variationnels Wasserstein, a démontré que l'utilisation de la distance de Wasserstein pour régulariser l'espace latent offre une plus grande flexibilité et robustesse. En s'inspirant de ces travaux, nous proposons une méthode, détaillée dans Haberstich et al. (en préparation), utilisant la distance de Wasserstein pour contraindre l'espace latent dans le cadre la résolution de problème inverse :

$$\hat{\mathbf{x}}^{\mathrm{W}} = G_\theta(\hat{\mathbf{z}}^{\mathrm{W}}) \quad | \quad \hat{\mathbf{z}}^{\mathrm{W}} \in \arg\min_{\mathbf{z}\in\mathbb{R}^k} \|\mathcal{A}(G_\theta(\mathbf{z})) - \mathbf{y}\|^2 + \lambda_{\mathrm{W}} \mathrm{W}_2^2(p(\mathbf{z}), \mathcal{N}(0, I_k)), \tag{8}$$

où λ_{W} est un paramètre de régularisation et $\mathrm{W}_2^2(p(\mathbf{z}), \mathcal{N}(0, I_k))$ définie par :

$$W_2^2(p(\mathbf{z}), \mathcal{N}(0, I_k))) = \inf_{\gamma\in\Pi(p(\mathbf{z}), \mathcal{N}(0,I_k))} \mathbb{E}_{(\mathbf{Z},\mathbf{U})\sim\gamma} \left[(\mathbf{Z} - \mathbf{U})^2\right],$$

où $\mathbf{Z}$ et $\mathbf{U}$ sont des variables aléatoires de loi respectives $p(\mathbf{z})$ et $\mathcal{N}(0, I_k)$, $\Pi(p(z), \mathcal{N}(0, I_k))$ est l'ensemble des distributions de probabilité conjointes $\gamma(\mathbf{Z}, \mathbf{U})$ ayant $p(\mathbf{z})$ et $\mathcal{N}(0, I_k)$ comme marginales.

En pratique, l'optimisation peut se faire avec un algorithme de gradient (par exemple, descente de gradient stochastique ou Adam) pour trouver le vecteur latent optimal tout en respectant la régularisation. Il n'est dès lors pas possible de calculer $\mathrm{W}_2^2(p(\mathbf{z}), \mathcal{N}(0, I_k))$ car on ne dispose que d'une réalisation unique du vecteur latent $\mathbf{z}^i = (z_1^i, z_2^i, \ldots, z_k^i)$ à chaque itération. Pour contourner cette difficulté, on considère chaque coordonnée z_j^i $(j = 1, \ldots, k)$ de $\mathbf{z}^i$ comme une réalisation indépendante d'une variable gaussienne $\mathcal{N}(0, 1)$. On cherche alors à mesurer la distance de Wasserstein entre la distribution gaussienne centrée réduite et la distribution empirique associée à $\mathbf{z}^i$ définie par :

$$p_{\text{emp}}(\mathbf{z}^i) = \frac{1}{k} \sum_{j=1}^{k} \delta_{z_j^i}, \tag{9}$$

où $\delta_{z_j^i}$ est la mesure de Dirac centrée en z_j^i. Le problème inverse devient :

$$\hat{\mathbf{x}}^{\mathrm{We}} = G_\theta(\hat{\mathbf{z}}^{\mathrm{We}}) \quad | \quad \hat{\mathbf{z}}^{\mathrm{We}} \in \arg\min_{\mathbf{z}\in\mathbb{R}^k} \|\mathcal{A}(G_\theta(\mathbf{z})) - \mathbf{y}\|^2 + \lambda_{\mathrm{We}} \mathrm{W}_2^2(p_{\text{emp}}(\mathbf{z}^i), \mathcal{N}(0, 1)), \tag{10}$$

où λ_{We} est un paramètre de régularisation. De manière générale, le calcul exact de calcul de la distance de Wasserstein est complexe voire impossible, mais dans le cas unidimensionnel entre une distribution empirique et une gaussienne centrée réduite, il existe des formules explicites, que l'on trouve notamment dans Peyré et Cuturi (2019) basées sur les moments, les quantiles et les moments conditionnels.

5 Application aux expériences laser

Le principe d'un laser de puissance tel que le Laser MegaJoule est de chauffer de la matière à très haute température (plusieurs millions de degrés) afin de crée un plasma (gaz dense et chaud) qui va ré-émettre un rayonnement X intense que l'on cherchera à caractériser ou à utiliser comme source d'énergie. Une mesure expérimentale consiste à obtenir une image des rayons X à un instant donné. L'objectif de ce travail est de déterminer les cartes de température et de densité des matériaux décrivant l'état de la matière à partir de ces images X, illustrées sur la figure 1.

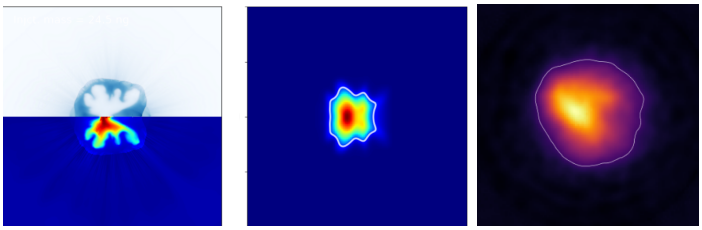


FIGURE 1 – A gauche : cartes de densité (en haut) et de températures (en bas) - au centre : image X simulée - à droite : image X expérimentale.

Il est nécessaire de construire la base de données pour l'entraînement du WGAN-GP, cela nécessite plusieurs milliers d'images. Les simulations laser sont très coûteuses, il nous est impossible d'obtenir un tel nombre d'images à l'heure actuelle. Nous avons recours à un code de physique approchée (modélisation du point chaud) dans lequel nous faisons varier des paramètres physiques. Nous avons ainsi obtenu 5000 cartes de température représentatives de la physique mise en jeux. La figure 2 illustre les résultats obtenus. On observe que seulement la régularisation de Wasserstein permet de retrouver une image sans "trous".

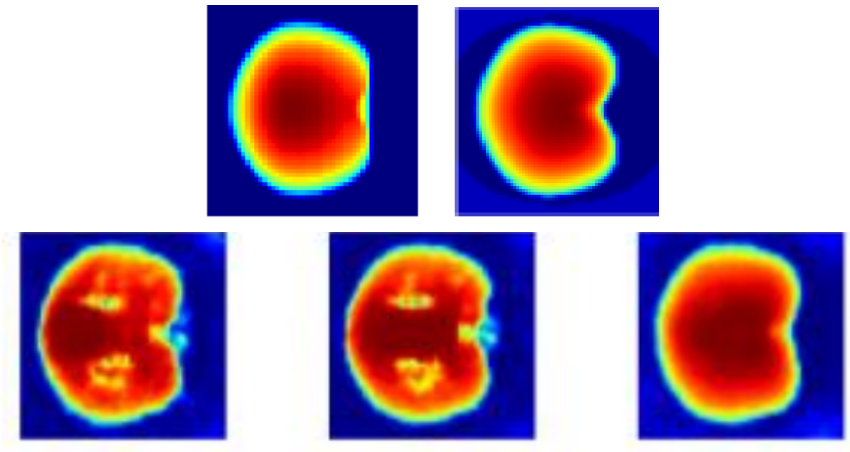


FIGURE 2 – Haut : la carte de température (gauche) et l'image X(droite) à retrouver – Bas : résultats obtenus avec trois types de régularisation dans l'espace latent (Clipping, L_2 et Wasserstein).

Bibliographie

Arjovsky, M., Chintala, S. & Bottou, L. (2017), Wasserstein Generative Adversarial Networks. Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pp. 214-223.

Arridge, S. R. (2019), Computed Tomography : Principles, Techniques, and Applications. John Wiley & Sons.

Dumoulin, V., Manzagol, P., Courville, A. & Bengio, Y. (2018), Generative Inverse Problems with Generative Adversarial Networks. In Advances in Neural Information Processing Systems, 31.

Feizi, S., Liao, Y. & Barahman, M. M. (2018), Diffusion Models for Non-Equilibrium Systems. Proceedings of the 35th International Conference on Machine Learning, pp. 1041-1050.

Goodfellow, I. et al. (2014), Generative adversarial networks. Advances in Neural Information Processing Systems, 27, pp. 2672-2680. Disponible à l'adresse :

Gulrajani, I., Ahmed, F., Doersch, C. & Lakshminarayanan, B. (2017), Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems, 30, pp. 5767-5777.

Habertich, C., Galerne, B. & Abraham, I. (en préparation), Latente Wassertstein regularization for GAN-based imaging inverse problems.

Kingma, D. P. & Welling, M. (2014), Auto-Encoding Variational Bayes. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 1250-1258.

Ongie, M., Bresler, Y. & Schniter, P. (2020), Deep learning techniques for inverse problems in imaging. IEEE Journal on Selected Areas in Information, 18(3), pp. 528-542.

Paris, S., Protter, N. & Mallat, S. (2017), Generative Adversarial Networks for Inverse Problems in Imaging. IEEE Transactions on Image Processing, 26(10), pp. 4632-4644.

Peyre, G. & Cuturi, M. (2019), Computational Optimal Transport : With Applications to Data Science. Foundations and Trends in Machine Learning, 11(5-6), pp. 355-607.

Rudin, L., Osher, S. & Fatemi, E. (1992), Nonlinear total variation based noise removal algorithms. Physica D : Nonlinear Phenomena, 60(1-4), pp. 259-268.

Santambrogio, F. (2015), Optimal Transport for Applied Mathematicians : Calculus of Variations, PDEs and Modeling. Birkhauser.

Tikhonov, A. N. (1963), Solution of incorrectly formulated problems and the regularization method. Sov. Math. Dokl., English translation of Doklady Akademii Nauk SSSR, 4(1), pp. 1035-1038.

Tolstikhin, I., Bkolb, N., Starchenko, F. & Vetrov, P. (2018), Wasserstein Auto-Encoders. Journal of Machine Learning Research, 19(71), pp. 1-31.

Problèmes inverses et apprentissage informé par la physique

Hugo Gangloff [1] & Nicolas Jouvin [2]

[1] *INRAE, France et hugo.gangloff@inrae.fr*
[2] *INRAE, France et nicolas.jouvin@inrae.fr*

Résumé. Dans le contexte des réseaux de neurones informés par la physiques (PINNs), nous cherchons à estimer les paramètres d'un système dynamique observé de manière bruitée et/ou indirecte. Ce problème inverse est présent dans de nombreuses applications, de la physique à la biologie ou l'écologie. Nous présentons une méthodologie générale d'apprentissage basée sur la minimisation d'une fonction de coût incorporant un terme d'attache aux données avec une régularisation issue de la physique. Nous illustrons l'approche dans deux contextes différents d'estimation de paramètres : pour un système d'équations aux dérivées ordinaires, et pour une équation aux dérivés partielles (EDP) de type Fokker-Planck. Nous mettons en avant de premiers résultats numériques motivant notre approche.

Mots-clés. PINNs, Problème inverse, Régularisation

Abstract. In the context of Physics-Informed Neural Networks (PINNs), we aim to estimate the parameters of a partially observed dynamic system corrupted by noise. This inverse problem is encountered in many applications, from physics to biology or ecology. We present a general learning methodology based on the minimization of a cost function integrating a data-fitting term with a physical regularization. We illustrate our approach in two different parameter estimation contexts : for a system of ordinary differential eqution, and for a Fokker-Planck type partial differential equation. We show promising preliminary numerical results motivating our approach.

Keywords. Physics-informed neural networks, Inverse problems, Regularization

1 Introduction

Depuis l'article fondateur de Raissi et al. (2019), les réseaux de neurones informés par la physique (PINNs) se sont imposés comme un sujet de recherche majeur. Permettant d'injecter de l'a priori « physique » dans l'apprentissage, les PINNs peuvent être vus comme une méthode d'apprentissage régularisée par une équation différentielle, ou comme une alternative aux solveurs numériques (Cuomo et al. 2022). Les recherches actuelles vont de l'amélioration de l'apprentissage (Wu et al. 2023), aux méta-modèles (Avila Belbute-Peres et al. 2021) et à l'analyse théorique (Doumèche et al. 2023), en passant par l'élaboration de benchmarks d'évaluation communs (Hao et al. 2023).

Les problèmes inverses sont encore relativement peu étudiés dans la littérature PINN. dans cette étude nous cherchons à estimer les paramètres $\theta \in \mathbb{R}^p$ d'un système dynamique observé de manière bruitée et/ou indirecte. Dans la Section 2, nous introduisons une méthodologie basée sur les PINNs. La Section 3 l'illustre dans deux cas d'études, celui d'un système de Lotka-Volterra généralisé en microbiologie, et celui d'une EDP de Fokker-Planck issue d'une équation différentielle stochastique.

2 Fonction de coût régularisée par la physique et estimation de paramètres

Dans le cadre de l'apprentissage informé par la physique, on dispose d'une équation différentielle décrite par un opérateur différentiel $\mathcal{N}_\theta$ - où θ représente les paramètres - qui définit la "physique" du problème. Dans la suite, nous exposons le cas des EDPs où l'on cherche la fonction u solution de :

$$\forall t \in I, \forall x \in \Omega, \quad \mathcal{N}_\theta[u](t,x) = 0, \tag{1}$$

où I est un intervalle de temps et $\Omega \subset \mathbb{R}^d$ est un domaine spatial. Notons que le cadre s'étend naturellement au cas stationnaire et aux systèmes d'EDOs, où u ne dépend pas de t ou x, respectivement. L'opérateur $\mathcal{N}_\theta$ peut contenir des dérivées partielles et de différents ordres selon le problème considéré, par exemple l'équation de la chaleur est définie par $\mathcal{N}_\theta = \partial_t - \theta\Delta$ avec $\theta > 0$, le paramètre de diffusion. Des conditions initiales et/ou de bord peuvent également être fixées.

On peut alors définir les résidus de l'équation, évalués sur les *points de collocations* $\{t_i, x_i\}_{i=1}^{n_c}$, généralement tirés au hasard sur $I \times \Omega$:

$$\|\mathcal{N}_\theta[u]\|^2 \coloneqq \frac{1}{n_c} \sum_{i=1}^{n_c} |\mathcal{N}_\theta[u](t_i, x_i)|^2. \tag{2}$$

De la même manière, on peut définir des termes de résidus $\|u - u_0\|_{IC}^2$ et $\|\mathcal{B}[u]\|_{BC}^2$ pour les éventuelles conditions initiales et de bord, respectivement. L'idée des PINNs est de chercher la solution u dans une famille de fonctions paramétriques $\mathcal{F} = \{u_\nu,\ \nu \in \mathbb{R}^k\}$ en minimisant les résidus via la fonction de coût suivante :

$$\mathcal{L}_{PINN}(\theta, \nu) = \|\mathcal{N}_\theta[u_\nu]\|^2 + \|u - u_0\|_{IC}^2 + \|\mathcal{B}[u]\|_{BC}^2. \tag{3}$$

La famille de fonctions $\mathcal{F}$ est choisie comme une classe de réseaux de neurones, où ν représente l'ensemble des poids et biais des différentes couches. Ce choix est motivé à la fois par leur bonnes capacités d'approximation (Cybenko 1989), ainsi que leur adéquation avec la différentiation automatique (Baydin et al. 2018), qui permet d'évaluer de manière exacte et peu coûteuse l'opérateur $\mathcal{N}_\theta[u_\nu](t_i, x_i)$ sur les points de collocation.

Lorsque que les paramètres de l'équation θ sont connus, minimiser $\mathcal{L}_{PINN}(\nu)$ consiste à approcher la solution de l'équation, comme le font les schémas numériques type éléments ou volumes finis. Dans le cas où l'on dispose d'observations directes et possiblement bruitées de

la solution $\mathcal{D} = \{(t_j, x_j, u_j)\}_{j=1}^n$, on peut rajouter un terme d'attache aux données de type moindres carrés,

$$\mathcal{L}_{\mathcal{D}}(u_\nu) = \frac{1}{n}\sum_{j=1}^{n} |u_\nu(t_j, x_j) - u_j|^2. \tag{4}$$

La minimisation de la fonction de coût globale $\mathcal{L} := \mathcal{L}_{PINN} + \mathcal{L}_{\mathcal{D}}$ peut alors être vue comme un problème de régression *régularisé* par la physique, où l'on cherche à interpoler les données $\mathcal{D}$ tout en respectant la dynamique prescrite par l'équation différentielle. Dans ce travail, on s'intéresse également à des termes d'attache aux données $\mathcal{L}_{\mathcal{D}}(u_\nu)$ plus généraux que les moindres carrés, où la solution u n'est pas directement observée. Par exemple lorsque l'on pose un modèle statistique sur le bruit d'observations des données $\mathcal{D}$ conditionnellement à la solution u de l'EDP : $\mathcal{D} \mid u \sim p(\cdot \mid u)$, où la dépendance en u peut être complexe. Le terme d'attache aux données est alors défini comme (moins) la log-vraisemblance des observations $\mathcal{L}_{\mathcal{D}}(u_\nu) = -\log p(\mathcal{D} \mid u_\nu)$. La régression avec l'Équation (4) en est un cas particulier où l'on dispose d'observations directes et indépendantes de la solution en des points (t_j, x_j), avec un bruit de mesure Gaussien $u_j \mid u \sim \mathcal{N}(\cdot \mid u(t_j, x_j), 1)$. Des modèles complexes du bruit d'observation ont été proposés, par exemple en écologie spatiale pour modéliser le déplacement d'espèces via des modèles « mecanico-statistiques » (Soubeyrand et al. 2014) ou des diffusions stochastiques (Gloaguen et al. 2018).

De manière générale, le problème *inverse* s'intéresse à l'estimation des paramètres de l'équation θ à partir des observations $\mathcal{D}$. On suppose que le terme d'attache aux données est différentiable par rapport aux paramètres ν. C'est le cas pour la regression mais aussi pour la plupart des modèles. Dans la section suivante, nous illustrons deux méthodologies possibles pour l'estimation de θ.

1. **Méthode itérative :** il s'agit d'une montée de coordonnées sur le coût global en ν et θ alternativement :

$$\begin{aligned} \theta^{(k+1)} &\leftarrow \operatorname{argmin}_\theta \mathcal{L}_{PINN}(\theta, \nu^{(k)}), \\ \nu^{(k+1)} &\leftarrow \operatorname{argmin}_\nu \mathcal{L}_{PINN}(\theta^{(k+1)}, \nu) + \lambda \mathcal{L}_{\mathcal{D}}(u_\nu). \end{aligned} \tag{5}$$

   L'algorithme s'arrête après un nombre d'itérations prédéfini ou lorsque le gain en coût global atteint un certain seuil. Cet algorithme a été proposé sous le nom de Generalized Smoothing Algorithm (GSA) par Ramsay et al. (2007) dans le cadre de la régression pour les EDOs. Cet algorithme se base sur les splines plutôt que les réseaux de neurones comme famille paramétrique u_ν, les premières permettant un calcul analytique de $\mathcal{N}_\theta$.

2. **Optimisation directe** : cette méthode minimise directement le coût global par descente de gradient :

$$\arg\min_{\nu,\theta} \mathcal{L}(\nu, \theta). \tag{6}$$

Dans les deux cas, le choix de l'algorithme d'optimisation est libre, et l'optimisation stochastique peut-être utilisée lorsque l'on utilise un grand nombre de points de collocation n_c ou que le nombre d'observations est grand. Cette méthode permet donc d'estimer les paramètres θ, bien que la non-convexité en ν rende difficile l'analyse des propriétés théoriques des estimateurs.

3 Résultats préliminaires

Nous présentons deux expériences motivant notre approche, l'une dans un cadre de régression pour les EDOs, et l'autre dans un cadre d'estimation des paramètres d'une diffusion. Les expériences de cette section ont été réalisées avec la librairie Python `jinns` (Gangloff et al. 2024).

Estimation des paramètres d'un système de Lotka-Volterra généralisé On s'intéresse à l'estimation des paramètres dans un système de Lotka-Volterra Généralisé (GLV) à N_s populations. Nous renvoyons le lecteur vers Hossie et al. (2024) pour une mise en contexte biologique du problème et l'étude du problème *forward* avec θ fixé. Mathématiquement, le système GLV peut s'écrire de manière matricielle comme :

$$\frac{\partial}{\partial t}u(t) = \mu + Ae^{u(t)}, \quad t \in [0, t_{\max}]\,. \tag{7}$$

où $u(t) \in \mathbb{R}^{N_s}$ désigne la (log-)quantité des N_s populations au temps t, μ_i est la taux d'accroissement de la population i et $A_{i,j}$ mesure l'interaction entre les populations i et j.

Munis d'observations directes $\mathcal{D} = \{(t_j, u_j)\}_{j=1}^n$ et possiblement bruitées du système, nous souhaitons estimer les paramètres du système notés $\theta = \{\mu, A\}$ à l'aide de l'algorithme décrit dans (5), dénommé GSA-PINN, et la perte des moindres carrés (4). Ce problème a déjà été investigué par Brunel et al. (2018) en utilisant l'algorithme GSA basé sur les splines

Pour un nombre de populations N_s donné, on fixe des paramètres $\theta^\star$ et on génère des données synthétiques simulées selon le vrai système dynamique. On échantillonne à des temps réguliers 100 points u_i^j par population, soit $n = 100N_s$ observations, et ajoutons un bruit Gaussien d'écart-type $\sigma = 0,1$. L'optimisation en θ est faite via un algorithme de gradient proximal avec régularisation l^1 (programme convexe dans le cas présent), tandis que l'optimisation en ν est faite par descente de gradient. L'architecture de réseau de neurones utilisée est un perceptron à 5 couches avec $7N_s$ neurones par couche et activation tanh, ce qui est standard dans la littérature PINNs.

Le Tableau 1 donne l'erreur quadratique relative d'estimation pour $\hat{\theta}$ ainsi que le temps de calcul. Les estimateurs $\hat{\theta}$ obtenu avec GSA-PINN sont toujours meilleurs ou équivalents à ceux de l'algorithme GSA, avec des gains de temps non négligeables.

Estimation de la loi stationnaire et des paramètres d'une diffusion : l'exemple d'Ornstein-Uhlenbeck Dans ce deuxième exemple, on propose une preuve de concept de notre méthodologie pour apprendre la loi et les paramètres d'une équation différentielle stochastique (EDS). On s'intéresse ici au processus d'Ornstein-Uhlenbeck (OU) sur $\mathbb{R}$ définit par $\mathrm{d}X_t = \alpha(\mu - X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}B_t$ où B_t est le mouvement Brownien sur $\mathbb{R}$, $\mu \in \mathbb{R}$, $\alpha, \sigma > 0$. La loi stationnaire existe et vérifie l'équation de Fokker-Planck (Risken 1996) suivante :

$$\mathcal{N}_\theta[u](x) \coloneqq \partial_x(\mu - x)u_{\text{statio}}(x) + \gamma\partial_x^2 u_{\text{statio}}(x) = 0, \quad \text{où : } \gamma = \frac{\sigma^2}{2\alpha},$$

	GLV ($N_s = 3$)		GLV ($N_s = 20$)		OU (μ)		OU (γ)	
	GSA	GSA-PINN	GSA	GSA-PINN	Emp.	PINN	Emp.	PINN
$\|\hat{\theta} - \theta\|$	1.720	**1.195**	0.104	**0.101**	**1.18e-2**	2.44e-2	2.01e-2	**5.77e-3**
Temps (s)	56	**15**	357	**63**	**0.01**	44	**0.12**	44

TABLE 1 – Comparaison de la méthodologie PINN avec des alternatives pour l'estimation de paramètres d'un système GLV et d'un processus d'Ornstein-Uhlenbeck (OU). La ligne ($\|\hat{\theta} - \theta\|$) compare le vrai θ et l'estimateur via une RMSE. Pour OU, le temps de calcul des PINNs comprend l'estimation de la loi stationnaire via u_ν en plus de $\hat{\theta}$.

où les paramètres de l'équation sont $\theta = (\mu, \gamma)$. Le processus OU est un cas simple de diffusion où la loi stationnaire est une gaussienne $u_{\text{statio}} = \mathcal{N}(\mu, \frac{\sigma^2}{2\alpha})$. Dans ce cas précis, la méthodologie décrite dans la Section 2 permet d'estimer les paramètres θ et la loi stationnaire u_{statio} via un PINN u_ν. Supposons que l'on observe $\mathcal{D} = \{X_{t_j}\}_{j=1}^n$ une trajectoire du processus dans son état stationnaire à n temps discrets t_j. On considère la log-vraisemblance composite (en « cassant les dépendances ») des données sous u_ν :

$$\mathcal{L}_{\mathcal{D}}(u_\nu) = \frac{1}{n} \sum_{j=1}^{n} \log u_\nu(X_{t_j}).$$

Notons que dans ce cas précis, u_ν est une densité de probabilité et doit intégrer à 1. On peut rajouter une pénalisation $|\int u_\nu - 1|^2$ dans le coût total ou utiliser des architectures de réseaux qui respectent directement cette contrainte (*e.g.* Normalizing Flows KOBYZEV et al. 2020).

On simule la trajectoire d'un processus OU en régime stationnaire, avec $(\mu, \gamma) = (2, 0.20)$ et $n = 2000$ échantillons régulier. On utilise l'algorithme d'optimisation directe décrit dans l'Équation (6), avec l'algorithme Adam en partant d'une initialisation $\theta_{\text{init}} = (1.2, 0.40)$. La Table 1 montre l'erreur MSE sur θ, se comparant favorable avec l'erreur sur les estimateurs empiriques de μ et σ. Cette preuve de concept ouvre la voie à l'estimation de paramètres pour des diffusions plus complexes $\mathrm{d}X_t = \mu_\theta(X_t)\,\mathrm{d}t + \sigma_\theta(X_t)\,\mathrm{d}B_t$, avec drift μ_θ et diffusion σ_θ paramétrés par θ, et dont la loi stationnaire n'est pas connue.

4 Conclusion

Ce travail propose une méthodologie générale pour les problèmes inverses consistant à estimer les paramètres d'une équation différentielle en fonction de données indirectes et bruitées. Les cadres applicatifs sont nombreux, notamment en sciences du vivant.

Remerciements Ces travaux sont réalisés pour partie en collaboration avec Lucia Clarotto, Sophie Donnet, Marie-Pierre Etienne, Yanis Jouanaud, Adeline Leclercq-Samson et Lorenzo Sala.

Références

Avila Belbute-Peres, Filipe de, Yi-fan Chen et Fei Sha (2021). "HyperPINN : Learning parameterized differential equations with physics-informed hypernetworks". In : *The symbiosis of deep learning and differential equations* 690 (cf. p. 1).

Baydin, Atilim Gunes et al. (2018). "Automatic differentiation in machine learning : a survey". In : *Journal of machine learning research* 18.153, p. 1-43 (cf. p. 2).

Brunel, Nicolas et al. (déc. 2018). "Parameter estimation for dynamical systems using an FDA approach". In : *11th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2018)*. Pise, Italy (cf. p. 4).

Cuomo, Salvatore et al. (2022). "Scientific machine learning through physics–informed neural networks : Where we are and what's next". In : *Journal of Scientific Computing* 92.3, p. 88 (cf. p. 1).

Cybenko, George (1989). "Approximation by superpositions of a sigmoidal function". In : *Mathematics of control, signals and systems* 2.4, p. 303-314 (cf. p. 2).

Doumèche, Nathan, Gérard Biau et Claire Boyer (2023). "Convergence and error analysis of PINNs". In : *arXiv preprint arXiv :2305.01240* (cf. p. 1).

Gangloff, Hugo et Nicolas Jouvin (2024). "jinns : a JAX Library for Physics-Informed Neural Networks". In : *arXiv preprint arXiv :2412.14132* (cf. p. 4).

Gloaguen, Pierre, Marie-Pierre Etienne et Sylvain Le Corff (2018). "Stochastic differential equation based on a multimodal potential to model movement data in ecology". In : *Journal of the Royal Statistical Society Series C : Applied Statistics* 67.3, p. 599-619 (cf. p. 3).

Hao, Zhongkai et al. (2023). "Pinnacle : A comprehensive benchmark of physics-informed neural networks for solving pdes". In : *arXiv preprint arXiv :2306.08827* (cf. p. 1).

Hossie, Paguiel Javan et al. (fév. 2024). "Simulating interactions in microbial communities through Physics Informed Neural Networks : towards interaction estimation". In : working paper or preprint (cf. p. 4).

Kobyzev, Ivan, Simon JD Prince et Marcus A Brubaker (2020). "Normalizing flows : An introduction and review of current methods". In : *IEEE transactions on pattern analysis and machine intelligence* 43.11, p. 3964-3979 (cf. p. 5).

Raissi, Maziar, Paris Perdikaris et George E Karniadakis (2019). "Physics-informed neural networks : A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". In : *Journal of Computational physics* 378, p. 686-707 (cf. p. 1).

Ramsay, Jim O et al. (2007). "Parameter estimation for differential equations : a generalized smoothing approach". In : *Journal of the Royal Statistical Society Series B : Statistical Methodology* 69.5, p. 741-796 (cf. p. 3).

Risken, H (1996). *The Fokker-Planck Equation* (cf. p. 4).

Soubeyrand, Samuel et Lionel Roques (2014). "Parameter estimation for reaction-diffusion models of biological invasions". In : *Population ecology* 56, p. 427-434 (cf. p. 3).

Wu, Chenxi et al. (2023). "A comprehensive study of non-adaptive and residual-based adaptive sampling for physics-informed neural networks". In : *Computer Methods in Applied Mechanics and Engineering* 403, p. 115671 (cf. p. 1).

PRÉDICTION DE CHAMPS PHYSIQUES SOUS CONTRAINTES LINÉAIRES

Nassouradine Mahamat[1,2], Clément Gauchy[1], Sébastien Da Veiga[2] & Pierre-Emmanuel Angeli[3]

[1] *Université Paris-Saclay, CEA, Service de Génie Logiciel pour la Simulation, France ; mahamat-hamdan.nassouradine@cea.fr, clement.gauchy@cea.fr*
[2] *ENSAI, CREST, F-35000 Rennes, France ; sebastien.da-veiga@ensai.fr*
[3] *Université Paris-Saclay, CEA, Service de Thermohydraulique et de Mécaniques des Fluides, France; pierre-emmanuel.angeli@cea.fr*

Résumé. La métamodélisation est une méthode essentielle en quantification d'incertitudes et optimisation des designs, car elle permet d'approximer des simulations coûteuses pour réaliser un grand nombre de prédictions à moindre coût. Les sorties de ces simulations peuvent être très complexes, comme des champs spatiaux sur des maillages fins ou des champs temporels. Elles peuvent également inclure plusieurs champs corrélés, parfois soumis à des contraintes linéaires. Les techniques existantes, comme la régression par processus Gaussiens combinée à des méthodes de réduction de dimension telles que l'ACP, sont efficaces pour gérer la grande dimensionnalité des données. Cependant, elles ne garantissent pas toujours le respect des contraintes dans les prédictions. Dans ce travail, nous proposons une méthode en deux étapes : une modélisation conjointe de plusieurs champs spatiaux à l'aide de processus Gaussiens multi-sorties (MOGP) combinés à l'ACP, puis une extension de cette méthode pour intégrer explicitement des contraintes linéaires entre les champs lorsque celles-ci sont connues. Enfin, nous présenterons un cas d'application de cette méthode sur un problème de dynamique des populations, basé sur le système d'équations de Lotka-Volterra.

Mots-clés. Processus Gaussien, ACP, Contraintes linéaires.

Abstract. Metamodeling is a key technique in uncertainty quantification and design optimization, enabling the approximation of expensive simulations to perform a large number of predictions at low cost. Simulation outputs are often complex, such as spatial fields on fine meshes or temporal fields, and may involve multiple correlated fields that are sometimes subject to linear constraints. Existing techniques, such as Gaussian process regression combined with dimensionality reduction methods like PCA, are effective for handling high-dimensional data. However, they do not always ensure that constraints are satisfied in predictions. In this work, we propose a two-step approach: first, a joint modeling of multiple spatial fields using multi-output Gaussian processes combined with PCA; second, an extension to account for linear constraints between fields when these constraints are known a priori. Finally, we will present an application of this method to a population dynamics problem derived from the Lotka-Volterra system of equations.

Keywords. Multi-Output Gaussian Process, PCA, linear constraint.

1 Introduction

Predicting physical fields is a high-dimensional regression problem that has gained significant interest in metamodeling for engineering applications. Gaussian Process (GP) [6] regression is widely used due to its probabilistic nature, which allows for the integration of prior information and provides uncertainty estimates. Extensions to multi-output GPs [8] have been developed to capture correlations between outputs and improve predictions. However, they are impractical for very high-dimensional outputs. As an alternative, [2] proposed combining PCA for output dimensionality reduction with independent GP modeling of latent weights, proving to be effective. Additionally, certain physical problems are governed by constraints, such as incompressibility in fluid mechanics or conservation laws, which can serve as prior knowledge to enhance models. In this context, the references [5, 1] introduced approaches to incorporate linear constraints into GPs by directly encoding them into the kernel. In this work, we focus on the simultaneous modeling of multiple high-dimensional scalar fields under linear constraints, leveraging a combination of PCA and constrained Gaussian Process regression in a two-step approach : a method for simultaneously modeling multiple scalar fields by combining PCA and multi-output Gaussian processes to flexibly capture potential dependencies between outputs. - an explicit integration of linear constraints into this model to ensure their satisfaction across different outputs.

2 Problem statement

The simulatneous modeling of multiple dependent scalar fields $u^j(\nu, \mathbf{x})$, $j = 1, \dots Q$, where ν denotes the spatial variable and $\mathbf{x} \in \mathbb{R}^D$ is a vector of parameters, generated from numerical simulations of a parametrized system of partial differential equations in a fixed mesh, can be considered as the approximation of a maping $\mathbf{f} = [\mathbf{f}^{1\top}, \dots, \mathbf{f}^{Q\top}] : \mathcal{X} \to \mathcal{M}$, where $\mathcal{X} \subset \mathbb{R}^D$ is the input space, and $\mathcal{M} \subset \mathbb{R}^P$ the corresponding output space. That is, $\mathbf{f}^j(\mathbf{x}) = \mathbf{y}^j = (u^j(\nu_1, \mathbf{x}), \dots, u^j(\nu_S, \mathbf{x}))$, $j = 1, \dots, Q$, $P = Q \times S$ and S is the mesh size. The objective of statistical emulation is to approximate the mapping $\mathbf{f}$ for a given set of training points from N simulations $(\mathbf{x}_i, \mathbf{y}_i = [\mathbf{y}_i^{1\top}, \dots, \mathbf{y}_i^{Q\top}])_{1 \le i \le N}$ such that $\mathbf{y}_i = \mathbf{f}(\mathbf{x}_i) \in \mathcal{M}, i = 1, ..., N$. In this work, we aim to incorporate linear constraints on the high dimensional output $\mathbf{f}(\cdot)$. Formally, we define this class of constraints as

$$\mathcal{F}[\mathbf{f}(\mathbf{x})] = \sum_{j=1}^{Q} \alpha^j \mathbf{f}^j(\mathbf{x}) = \mathbf{c}(\mathbf{x}) \tag{1}$$

where α^j are scalar coefficients and $c(.)$ a given function that we suppose perfectly known. This models linear relationships between components of $\mathbf{f}$ and which evolve according to the input parameters. For instance, we consider in this paper a Lokta-Volterra model according to [7] used in ecology for populations dynamics. Given $x(t)$ and $y(t)$ respectivly the population density of some prey and some of its predator, the population change through time follows the following system of differential equations:

$$\begin{cases} x'(t) = ax(t) - bx(t)y(t) \\ y'(t) = -cy(t) + dx(t)y(t) \end{cases}, \tag{2}$$

with a, b, c, d are the Lokta-Volterra model parameters. These prey-predator equations have the property of admitting Hamiltonian structure with respect to a generalized Poisson bracket as shown in [4] :

$$H(x, y) = dx - c\log(x) + by - a\log(y) \text{ such that } \frac{d}{dt}H(x(t), y(t)) = 0\ . \tag{3}$$

Equation (3) translates a non linear constraint on (x, y). We will illustrate the proposed model by predicting the evolution of the population (x, y) under a linearized Hamiltonian structure:

$$dx - cs + by - ar = c_0,$$

where $s = \log(x)$ and $r = \log(y)$. The model simultaneously predicts four outputs $(x(t, \eta), y(t, \eta), r(t, \eta), s(t, \eta))$ at each point of a predefined time discretization, with respect to the input parameter $\eta = (x_0, y_0)$, representing the initial population states. In the following, we will revisit multi-output GP theory before addressing output dimensionality by combining MOGP with principal component Analysis (PCA).

2.1 Gaussian process regression

A Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution. Formally, we write $f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, where $\mu(.)$ is the mean function and $k(\cdot, \cdot)$ the covariance function. Given a set of training points $(\mathbf{x}_i, z_i)_{1\leq i\leq N} \subset \mathbb{R}^D \times \mathbb{R}$ with matrix notation $(\mathbf{X}, \mathbf{Z})$, Gaussian Process Regression (GPR) models the relationship between inputs $\mathbf{x}_i \in \mathbb{R}^D$ and outputs $z_i \in \mathbb{R}$ as $z_i = f(\mathbf{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, where f is treated as a Gaussian process prior $f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ according to [6].

Multi-output Gaussian process : In the multi-output regression task, we aim to learn a vector-valued function $\mathbf{f} : \mathcal{X} \subset \mathbb{R}^D \to \mathbb{R}^L$ from training data $(\mathbf{x}_i, \mathbf{z}_i)_{1\leq i\leq N}$, where $\mathbf{z}_i = [z_i^1, \ldots, z_i^L]^\top$. Multi-output Gaussian Process (MOGP) regression approximates all L outputs $f_1, \ldots, f_L$ simultaneously, capturing their correlations to outperform independent modeling. A Gaussian process prior is placed on $\mathbf{f}(\mathbf{x}) \sim \mathcal{GP}(\boldsymbol{\mu}(\mathbf{x}), \mathbf{k}(\mathbf{x}, \mathbf{x}'))$, where $\boldsymbol{\mu}(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^L$ is the vector-valued mean function, and $\mathbf{k}(\mathbf{x}, \mathbf{x}') \in \mathbb{R}^{L\times L}$ is the matrix-valued covariance kernel as defined in [8] :

$$\mathbf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{L\times L}, \qquad \mathbf{k}(\mathbf{x}, \mathbf{x}') = \text{Cov}_{\mathbf{f}}[\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}')], \qquad [\mathbf{k}(\mathbf{x}, \mathbf{x}')]_{ll'} = \text{Cov}[f_l(\mathbf{x}), f_{l'}(\mathbf{x}')]$$

Popular kernel models for MOGPs include the Linear Model of Coregionalization (LCM) and its special case, the Intrinsic Coregionalization Model (ICM) exist in the literature. We refer to [8] for further details. In practice, the model's hyperparameters are optimized by maximizing the marginal log-likelihood. Gradient-based optimization methods are used to compute the maximum likelihood estimates of θ which parametrized the kernel. While MOGPs excel at capturing output correlations, their computational complexity $\mathcal{O}(N^3L^3)$ poses challenges for high-dimensional outputs.

2.2 PCA-MOGP model

To address the high dimensionality of output, a PCA is used to project the output into a lower-dimensional latent space in [2], followed by independent Gaussian process modeling of the latent

Algorithm 1: PCA-MOGP for simultaneous modeling of high dimensional vector field

Train :

Data : input : $\mathbf{X} \in \mathbb{R}^{N \times D}$

output : $\mathbf{Y}^1, \ldots, \mathbf{Y}^Q \in \mathbb{R}^{N \times S}$, Q scalar fields data from N simulation computed in a fixed mesh of size S

PCA : $\mathbf{Y} = [\mathbf{Y}^{1^\top}, \ldots, \mathbf{Y}^{Q^\top}]^\top \in \mathbb{R}^{NQ \times S}$

$\mathbf{H} = [\mathbf{v}_0^\top, \cdots, \mathbf{v}_M^\top] \leftarrow \text{PCA}(\mathbf{Y})$

Projection : $\mathbf{w}_i^j \leftarrow \mathbf{y}_i^{j^\top} \mathbf{H}$ for all $i = 1 : N$, $j = 1 : Q$

for $l = 1 : M$

MOGP : Train data : $(\mathbf{x}_i, \mathbf{w}_{i,l} = (w_{i,l}^1, \ldots, w_{i,l}^Q))_{1 \leq i \leq N}$

$\mathbf{w}_l \leftarrow \mathbf{h}_l(\mathbf{x})$, MOGP prior on $\mathbf{h}$

end for

Prediction : $\mathbf{w}_{*l} \leftarrow \mathbf{h}(\mathbf{x}_*)$ for all $l = 1 : M$

Reconstruction : $\mathbf{y}_*^j = \sum_{l=1}^M \mathbf{v}_l w_*^{l,j}$

variables. The simultaneous modeling of multiple spatial fields is based on the idea of simultaneously reducing the mesh size of all fields without taking into account the covariance between the fields at a same mesh point. PCA is applied in all the output data by considering the row-wise concatenation of the data derived from N simulations of each scalar field component. Then we model by MOGP simultaneously the M latent variables of the different fields to take into account their correlations, where M is the number of latent variables chosen by the elbow heuristic. The approach is summarized by Algorithm 1.

3 The PCA-MOGP model under linear constraint

To incorporate constraints into the previous model, the idea is to ensure that constraints are preserved during the dimensionality reduction step with PCA and effectively handled during modeling using constrained Gaussian processes in the MOGP framework, following a recent approach illustrated in [1]. Preserving constraints in PCA involves two steps: ensuring the latent variables satisfy the constraints derived from the constrained data and ensuring that the reconstructed data satisfy the constraints derived from the constrained latent variables. This can be achieved when the data are centered and the sum of the constraints is zero :

$$\begin{aligned} \mathbf{UY} = 0_{\mathbb{R}^{N \times S}} \rightarrow \mathbf{UW} = \mathbf{UYH}_M = 0_{\mathbb{R}^{R \times M}} \\ \mathbf{UW} = 0_{\mathbb{R}^{R \times M}} \rightarrow \mathbf{U\tilde{Y}} = \mathbf{UWH}_M^T = 0_{\mathbb{R}^{R \times S}} \end{aligned} \tag{4}$$

for any given constraint matrix $\mathbf{U}$, with $\mathbf{H}_M$ the reduced eigenvectors basis. To handle the case where the second term is non-zero, one way to do it is to transform the data to obtain a zero second term. For the case where the data satisfies:

$$\sum_j^Q \alpha^j \mathbf{y}_i^j = \mathbf{c}(\mathbf{x}_i), \quad i = 1, ..., N \equiv \sum_j^Q \alpha^j \mathbf{Y}^j = \mathbf{C}_x, \tag{5}$$

we can work with the transformation of one of the matrices $\mathbf{Y}^j$ defined by $\mathbf{Z} = \mathbf{Y}_1 - \alpha_1^{-1}\mathbf{C}_x$. The knowledge of $\mathbf{c}(x)$ for all $\mathbf{x} \in \mathcal{X}$ guarantees the uniqueness of the inverse transformation.

Constrained Gaussian process

Consider the problem of imposing a linear constraint on a vector-valued Gaussian process $\mathbf{f} \sim \mathcal{GP}(\boldsymbol{\mu}_f(\mathbf{x}), \mathbf{K}_f(\mathbf{x}, \mathbf{x}'))$, $\boldsymbol{\mu}_f(\cdot) : \mathbb{R}^D \mapsto \mathbb{R}^P, \mathbf{K}_f(\cdot, \cdot) : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{R}^P \times \mathbb{R}^P$ defined by the following linear operator:

$$\mathcal{L}[\mathbf{f}] = \mathbf{0} \tag{6}$$

where $\mathcal{L}$ is represented by the vector $\mathbf{u}_{gp} = [\alpha_1, ..., \alpha_Q]^\top$ such that

$$\mathcal{L}[\mathbf{f}] = 0 \iff \langle \mathbf{u}_{gp}, \mathbf{f}(\mathbf{x}) \rangle = \sum_{j=1}^{Q} \alpha^j f^j(\mathbf{x}) = 0, \quad \forall \mathbf{x} \in \mathcal{X} \tag{7}$$

The parametrization approach as described in [1] is based first on the principle of stability of Gaussian Processes under linear transformations, and then uses this principle to impose constraints on a linear operator, ensuring that the constraint is satisfied regardless of the underlying function. This technique involves considering that the function **f** is related to another function **g** via a linear operator $\mathcal{G}_x : \mathbb{R}^s \mapsto \mathbb{R}^P$ such that $\mathbf{f}(\mathbf{x}) = \mathcal{G}_x(\mathbf{g})$. The linear operator $\mathcal{G}_x$ is conditioned by the relation (6), such that $\mathcal{L}[\mathcal{G}_\mathbf{x}[\mathbf{g}]] = 0$ for any function **g**. Then a GP prior is placed on $\mathbf{g} \sim \mathcal{GP}(\boldsymbol{\mu}_g(\mathbf{x}), \mathbf{K}_g(\mathbf{x}, \mathbf{x}'))$, from which the prior on **f** is derived :

$$\boldsymbol{\mu}_f(\mathbf{x}) = \mathcal{G}_\mathbf{x}[\boldsymbol{\mu}_g] \qquad \mathbf{K}_f(\mathbf{x}, \mathbf{x}') = \mathcal{G}_\mathbf{x} \mathbf{K}_g \mathcal{G}_{\mathbf{x}'}^T \tag{8}$$

For example, given $\mathbf{f} = [f_1, f_2]$ which satisfies $f_1 + f_2 = 0$, $\mathcal{L}(\cdot)$ is represented by $\mathbf{u} = [1, 1]^\top$. The representative $\mathbf{B}$ of $\mathcal{G}_x$ is derived from $\mathcal{L}[\mathcal{G}_\mathbf{x}] = \mathbf{u}^\top \mathbf{B} = b_1 + b_2 = 0$. Therefore any vector of the form of $B = b[-1, 1]^\top$ parametrizes **f** which satisfies (6), and the explicit forms of the mean and the covariance function are given by

$$\boldsymbol{\mu}_h(\mathbf{x}) = \mathcal{G}_\mathbf{x}[\mu_g] = \begin{bmatrix} -\mu_g \\ \mu_g \end{bmatrix}, \qquad \mathbf{k}_h(\mathbf{x}, \mathbf{x}') = \mathcal{G}_\mathbf{x} k_g \mathcal{G}_{\mathbf{x}'}^T = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} k_g(\mathbf{x}, \mathbf{x}').$$

See [3] for further details on the algorithm for deriving the conditioned operator $\mathcal{G}_x$ in general case.

4 Application on a population dynamics problem

In order to assess the performance of PCA-MOGP under constraint, we focus on constructing a metamodel to simulate the Lotka-Volterra predator-prey equations. We aim to predict the population number (x, y) on a given temporal discretization $(t_1, \cdots, t_S)$ with $S = 1000$, based on the initial conditions (x_0, y_0). The data are generated by numerically solving the Lotka-Volterra system of equations using the Runge-Kutta method. This is done for 110 initial condition pairs (x_0, y_0), randomly sampled from the following distribution $(\mathcal{U}[1.9, 2.1], \mathcal{U}[0.28, 0.32])$ from which 10 initial conditions are kept for the training set and 100 for the test set. Fixed parameters $(a, b, c, d) = (1.1, 0.4, 0.4, 0.1)$ are used, and the same temporal discretization is applied to all solutions. We then considered the simultaneously modeling of $(\mathbf{x}, \mathbf{y}, \mathbf{s}, \mathbf{r}) \in \mathbb{R}^{4 \times S}$ subjected to the linear constraint $d\mathbf{x} - c\mathbf{s} + b\mathbf{y} - a\mathbf{r} = \mathbf{c}_0$. We considered two models for comparison to highlight the benefits of our approach : a constrained PCA-MOGP model (C-PCA-MOGP), an unconstrained

PCA-MOGP (PCA-MOGP) that models all four variables derived from the linear transformation of the non linear constraint. A PCA analysis of the training dataset has been performed and $M = 3$ latent variables have been chosen using the elbow heuristic for the 4 fields $(\mathbf{x}, \mathbf{y}, \mathbf{s}, \mathbf{r})$. Considering the testing set $(\hat{x}_i(t), \hat{y}_i(t), \hat{r}_i(t), \hat{s}_i(t))_{1 \leq i \leq n_{\text{test}}}$, we compare the two prediction models using different performance metrics, such that the testing L2 error:

$$e_{L2}(t) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} e_i(t) \ ,$$

where $e_i(t) = \sqrt{(x_i(t) - \hat{x}_i(t))^2 + (y_i(t) - \hat{y}_i(t))^2 + (r_i(t) - \hat{r}_i(t))^2 + (s_i(t) - \hat{s}_i(t))^2}$. Quantiles of the testing L2 error have also been estimated to account for the performances' variability in the testing set:

$$q_{\alpha,L2} = e_{(\lceil n_{\text{test}} \alpha \rceil)}(t) \ ,$$

where $\alpha \in (0, 1)$ is the quantile level. Performance with respect to the linear constraint is simply evaluated by

$$h(t) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} h_i(t) \ ,$$

where $h_i(t) = d\hat{x}_i(t) - c\hat{s}_i(t) + b\hat{y}_i(t) - a\hat{r}_i(t) - c_i(t)$, quantiles are also estimated in the same fashion than for the L2 loss:

$$q_{\alpha,\mathcal{C}} = h_{(\lceil n_{\text{test}} \alpha \rceil)}(t) \ ,$$

Let us remind that for the linear constraint to be satisfied we should obtain $h_i(t) = 0$ for $1 \leq i \leq n_{\text{test}}$ and $0 \leq t \leq T$. Figure 1 illustrates the performance of the two models both in terms of

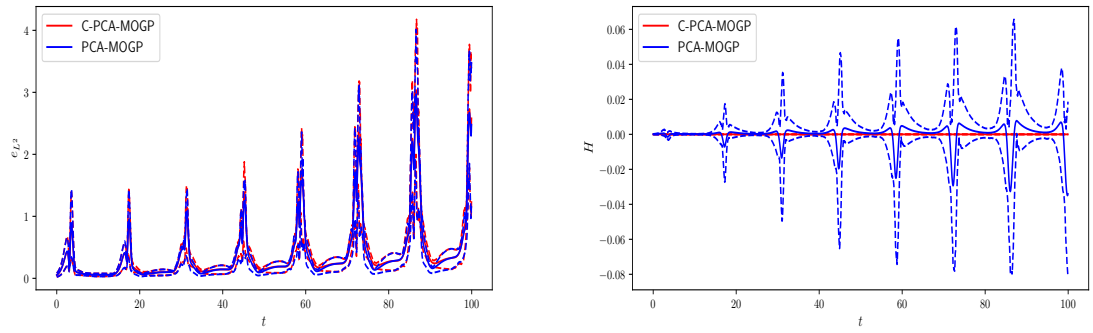


Figure 1: Left: Estimation of the L2 error evaluated on the testing dataset for the unconstrained PCA-MOGP (in blue) and constrained PCA-MOGP (in red). $q_{\alpha,L2}$ are estimated with $\alpha = 0.25$ and $\alpha = 0.75$ for both model (in dashed lines). Right: Estimation of the linear constraint on the testing dataset for the unconstrained PCA-MOGP (in blue) and constrained PCA-MOGP (in red). $q_{\alpha,\mathcal{C}}$ are estimated with $\alpha = 0.25$ and $\alpha = 0.75$ for both model (in dashed lines).

prediction using the L2 error and in terms of respect of the linear constraint. The reader can notice that the constrained PCA-MOGP model satisfy the constraint for all 100 initial conditions in the testing set and all time t, whereas the unconstrained PCA-MOGP model can violate the constraint up to 10^{-1}, while conserving basically the same performance in terms of L2 error.

5 Conclusion & perspectives

A methodology for predicting several high-dimensional physical fields under linear constraints using Gaussian process regression is illustrated in this article. First, the multi-output Gaussian process (MOGP) setting is introduced. Then, we demonstrate that a linear dimensionality reduction algorithm such that PCA allows to keep the linear constraints into the latent space, we use specific covariance kernels for the Gaussian process such that the linear constraint is satisfied both for prediction and samples from the posterior distribution of the Gaussian process. It assures that the linear constraint is strictly respected in the model's prediction, unlike popular physics-informed methods based on collocation points. The methodology proposed is benchmarked numerically on the famous Lotka-Volterra model used for population dynamics in ecology. An unconstrained and constrained PCA-MOGP regression model are tested and commented on this numerical application. The perspectives of this work is to implement this prediction algorithm in a two-dimensional fluid-dynamics computer simulation with eventually more complex constraints.

References

[1] A. WIlls et Thomas B.Shon Carl Jidling, Niklas Wahlstrom. Linearly constrained Gaussian processes. *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.*, pages 2–5, 2017.

[2] Dave Higdon, James Gattiker, Brian Williams, and Maria Rightley. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103(482):570–583, 2008.

[3] Markus Lange-Hegermann. Algorithmic Linearly Constrained Gaussian Processes. *Department of Electrical Engineering and Computer Science Ostwestfalen-Lippe University of Applied Sciences Lemgo, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.*, pages 2–5, 2018.

[4] Y Nutku. Hamiltonian structure of the lotka-volterra equations. *Physics Letters A*, 145(1):27–28, 1990.

[5] Philipp Pilar, Carl Jidling, Thomas B. Schön, and Niklas Wahlström. Incorporating sum constraints into multitask gaussian processes, 2022.

[6] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 11 2005.

[7] Peter J Wangersky. Lotka-volterra population models. *Annual Review of Ecology and Systematics*, 9:189–218, 1978.

[8] Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Modèles de dynamiques de population pour l'écologie et l'épidémiologie (Groupe Environnement) (Session du groupe)

Modèle mécanistico-statistique pour l'expansion de la maladie du flétrissement du frêne

Coralie Fritsch [1] & Marie Grodidier [2] & Anne Gégout-Petit [3] & Benoît Marçais [4]

[1] *Inria Nancy, IECL, France, coralie.fritsch@inria.fr*
[2] *INRAE Avignon, France, marie.grosdidier@inrae.fr*
[3] *Inria Nancy, IECL, France, anne.gegout-petit@univ-lorraine.fr*
[4] *INRAE Nancy, France, benoit.marcais@inrae.fr*

Résumé. *Hymenoscyphus fraxineus* est un champignon pathogène forestier invasif qui provoque un dépérissement sévère des populations de frênes européennes, appelé chalarose. La propagation de la maladie a été étroitement surveillée en France par le système d'enquête sur la santé des forêts. Nous avons développé un modèle mécanistico-statistique utilisant une équation aux dérivées partielles (EDP) qui décrit la propagation de la maladie. Il est particulièrement adapté à la dynamique des populations d'agents pathogènes induite par la propagation de spores et permet de prendre en compte le climat (températures estivales et précipitations printanières), l'effet Allee induit par des rencontres limitées avec des partenaires sexuels et la densité d'hôtes. Nous avons ajusté ce modèle à l'aide des données d'observation de la chalarose en France depuis 2008. Nous avons estimé les paramètres de notre modèle, en identifiant d'abord des gammes de valeurs appropriées pour les paramètres, ce qui a conduit à une réduction du modèle, puis en utilisant l'algorithme d'échantillonnage adaptatif AMIS (pour *adaptive multiple importance sampling*), pour l'ajustement. Le modèle reproduit bien la propagation observée en France au cours des 20 dernières années. Elle prévoit notamment l'absence d'impact de la maladie dans le sud-est du pays et son faible développement dans la vallée de la Garonne, dans le sud-ouest de la France. La température estivale est le facteur ayant l'effet global le plus important sur la propagation de la maladie et explique son impact limité dans le sud de la France. En revanche, l'effet Allee et l'hétérogénéité des précipitations printanières n'ont pas fortement affecté l'expansion globale de H. fraxineus en France et pourraient être négligés dans le processus de modélisation.

Mots-clés. Chalarose, maladie du flétrissement du frêne, pathogène invasif, équation de réaction-diffusion, algorithme AMIS, inférence Bayésienne.

Abstract. *Hymenoscyphus fraxineus* is an invasive forest fungal pathogen that induces severe dieback in European ash populations. The spread of the disease has been closely monitored in France by the forest health survey system. We have developed a mechanistic-statistical model using a partial differential equation (PDE) that describes the spread of the disease. It is particularly adapted for pathogen population dynamics driven by spore propagation and it allows us to take into account climate (summer temperature and spring rainfall), Allee effect induced by limited sexual partner encounters and host density. We fitted this model using available disease reports. We estimated the parameters of our model, first identifying the appropriate ranges for the parameters, which led to a model reduction, and then using an adaptive multiple importance sampling algorithm for fitting. The model

reproduces well the propagation observed in France over the last 20 years. In particular, it predicts the absence of disease impact in the south east of the country and its weak development in the Garonne valley in south-west France. Summer temperature is the factor with the highest overall effect on disease spread and explains the limited impact in southern France. In contrast, the Allee effect and the heterogeneity of spring precipitation did not strongly affect the overall expansion of H. fraxineus in France and could be neglected in the modeling process.

Keywords. *Chalara fraxinea*, Ash dieback, invasive pathogen, reaction-diffusion equation, AMIS algorithm, Bayesian inference.

1 Le modèle

La chalarose, ou maladie du flétrissement du frêne, est du à la propagation de l'agent pathogène *Hymenoscyphus fraxineus*. Nous avons développé un modèle mécanistico-statistique composé de trois sous-modèle : 1) un premier modèle de réaction-diffusion pour la propagation du pathogène, 2) un second modèle décrivant la probabilité de développement des symptômes de flétrissement des frênes en fonction de la quantité de pathogène, 3) et enfin un dernier modèle décrivant les données récoltées par le département de la santé des forêts.

1.1 Modèle de réaction-diffusion pour la dynamique du pathogène

Au printemps, le champignon *Hymenoscyphus fraxineus* présent dans les rachis infectés (feuilles infectées tombées aux pieds des arbres) produit des spores. Ces spores vont ensuite se disperser durant l'été sur les arbres voisins. Les feuilles des arbres qui ont été infectées par les spores vont, en tombant aux pieds des arbres à l'automne, constituer de nouveaux rachis infectés qui produiront des spores au printemps de l'année suivante. Le modèle de réaction-diffusion proposés ici décrit ces trois étapes.

Production de spores au printemps. Nous notons $R_a(x)$ la quantité de rachis infectés en la position x au printemps de l'année a. Nous décrivons la quantité $\nu_a(x)$ de spores produite à la position x durant le printemps de l'année a par

$$\nu_a(x) = \begin{cases} H_a(x)\, R_a(x) & \text{if } R_a(x) \geq r \\ \\ H_a(x)\, \frac{R_a(x)^2}{r} & \text{if } R_a(x) < r \end{cases} \tag{1}$$

où H_a dépend de la pluviométrie $h_a(x)$ en x au printemps de l'année a à travers la relation

$$H_a(x) = \beta_0 + \beta_1\, h_a(x). \tag{2}$$

Dans ce modèle, outre la prise en compte de l'humidité dans le développement du pathogène, nous supposons également qu'il est sujet à un effet Allee [7], décrit par le paramètre r ci-dessus.

Diffusion des spores durant l'été. On suppose que la quantité de spores $\nu_a(x)$ produite par les rachis est ensuite libérée uniformément et se diffuse durant une période de $\tau = 60$ jours durant l'été [2]. La dynamique de la quantité de spores $w_a(t,x)$ en x au temps t est alors donnée par l'équation de réaction-diffusion suivante

$$\frac{\partial w_a(t,x)}{\partial t} = D\,\Delta w_a(t,x) + \frac{\nu_a(x)}{\tau} \tag{3}$$

où $w_a(0,x) = 0$. À la fin de l'été, la répartition des spores est donnée par $w_a(\tau,\cdot)$.

Colonisation de nouveaux rachis durant l'automne. À l'automne, les feuilles infectées tombent, entraînant la production de rachis nouvellement colonisés sur le sol forestier. La quantité locale de feuilles infectées dépend à la fois de la densité des frênes $d(x)$ et du niveau d'infection des feuilles supposé être linéairement dépendant de la quantité de spores $w_a(\tau,x)$ à la fin de la période de diffusion. On note $\chi_a(x)$ la quantité de rachis nouvellement colonisés en x à l'automne de l'année a. On suppose qu'au-dessus d'un seuil de saturation S, l'augmentation de la charge en spores n'augmente pas l'infection car toutes les feuilles sont déjà infectées. Alors, $\chi_a(x)$ est donnée par

$$\chi_a(x) = (w_a(\tau,x) \wedge S)\, d(x). \tag{4}$$

Les rachis colonisés produisent des spores pendant plusieurs années [4, 6, 8]. Nous supposons que les rachis colonisés sont actifs durant deux ans. La quantité de rachis R_{a+1} est alors

$$R_{a+1}(x) = \chi_a(x) + \chi_{a-1}(x)\,. \tag{5}$$

1.2 Modèle pour la probabilité de développement des symptômes

Les symptômes de dépérissement des frênes sont causés par le transfert de l'agent pathogène des feuilles infectées vers les pousses avant la chute des feuilles. L'infection reste latente et devient perceptible durant l'hiver qui suit l'infection. Nous supposons qu'un frêne sans symptôme, en x, développe un dépérissement observable l'année a, avec une probabilité linéaire dépendante de la quantité d'agent pathogène infectant ses feuilles l'année $a-1$, i.e proportionnelle à $\chi_{a-1}(x)$. Les données dont nous disposons étant discrètes sur des quadrats, nous modélisons la probabilité de dépérissement sur ces quadrats. En supposant que les quantités d'agents pathogènes par arbre sont uniformes dans chaque quadrat, cette probabilité est proportionnelle à $\chi^i_{a-1}/d(i)$ où $\chi^i_a = \int_{\omega_i} \chi_a(x)\,\mathrm{d}x$ est la quantité totale d'agents pathogènes et $d(i)$ la densité de frênes dans le quadrat ω_i. De plus, le transfert des feuilles infectées vers les pousses sont entravés lorsque les températures estivales sont élevées ([5, 3, 9]). Finalement, un frêne du quadrat ω_i développe des symptômes observables l'année a avec une probabilité

$$\tilde{q}^i_a = \frac{\chi^i_{a-1} f(T^i_{a-1})}{r_S\, d(i)} \wedge 1, \tag{6}$$

où r_s est un paramètre de proportionnalité, T^i_a est un indice de température de l'été de l'année a et f est de la forme

$$f(T) = \left[1 - \frac{T}{\gamma}\right]^{\kappa}_{+}. \tag{7}$$

γ est le seuil de la variable de température à partir duquel le développement des symptômes n'est plus possible et $\kappa > 0$ décrit l'impact de valeurs faibles (mais positives) de T.

De plus, si un arbre développe des symptômes, ceux-ci persistent d'une année sur l'autre avec une probabilité que l'on note C_{pers}. Par conséquent, la probabilité qu'un arbre du quadrat ω_i présente des symptômes de dépérissement au cours de l'année a est

$$q_a^i = C_{\text{pers}}\, q_{a-1}^i + (1 - C_{\text{pers}}\, q_{a-1}^i)\, \tilde{q}_a^i\,. \tag{8}$$

Les probabilités initiales de symptômes de dépérissement sont dérivées des intensités de prévalence, connues à constante multiplicative C_{init} près.

1.3 Modèle statistique pour les données d'observation

Pour chaque année d'observation $a \in [\![2008, 2019]\!]$, I_a désigne l'ensemble des indices des quadrats observés au cours de l'année a et pour chaque $i \in I_a$, $\text{obs}_a(i)$ désigne le nombre de parcelles observées dans le quadrat ω_i au cours de l'année a. Les données observées sont données par $(p_a^k(i))_{a,i\in I_a,k\in[\![1,\text{obs}_a(i)]\!]}$ où $p_a^k(i)$ est la proportion d'arbres infectés pour la k-ième parcelle observée du quadrat ω_i dans l'année a. Sur chaque parcelle, environ $m = 30$ arbres sont observés de sorte que $m\,\text{obs}_a(i)$ est le nombre d'arbres observés dans le quadrat ω_i.

Nous supposons que chaque arbre est observé au plus une fois, de sorte que les différentes observations sont indépendantes. Par conséquent, le nombre N_a^i d'arbres infectés observés pour le quadrat ω_i dans l'année a est donné par le modèle binomial

$$N_a^i \sim \text{Bin}(m\,\text{obs}_a(i), q_a^i). \tag{9}$$

De plus, conditionnellement au processus $(q_a^i)_{a,i}$, les variables $(N_a^i)_{a,i}$ sont indépendantes.

Soit $p_a(i)$ la proportion observée d'arbres infectés dans le quadrat ω_i dans l'année a:

$$p_a(i) = \frac{1}{\text{obs}_a(i)} \sum_{k=1}^{\text{obs}_a(i)} p_a^k(i), \tag{10}$$

Le nombre $m\,\text{obs}_a(i)\,p_a(i)$ d'arbres infectés observés dans ω_i est une réalisation de N_a^i.

2 Inférence statistique

Nous utilisons une méthode d'inférence statistique bayésienne basée sur la vraisemblance. La vraisemblance d'un ensemble de paramètres $\theta = (D, \beta_0, \beta_1, r, \gamma, \kappa, S, C_{\text{init}}, r_S, C_{\text{pers}})$ est

$$Ł(\theta) = \prod_{a=2008}^{2019} \prod_{i\in I_a} \binom{m\,\text{obs}_a(i)}{m\;\text{obs}_a(i)\,p_a(i)} \left(q_a^i(\theta)\right)^{m\;\text{obs}_a(i)\,p_a(i)} \left(1 - q_a^i(\theta)\right)^{m\,\text{obs}_a(i)\,(1-p_a(i))} \tag{11}$$

où $(q_a^i(\theta))_{a,i}$ est la probabilité de symptômes de dépérissement défini par (8) avec l'ensemble de paramètres θ, et $p_a(i)$ est la proportion observée d'arbres infectés définie par (10).

Pour $\alpha > 0$, on remarque que les paramètres $\theta = (D, \beta_0, \beta_1, r, \gamma, \kappa, S, C_{\text{init}}, r_S, C_{\text{pers}})$ et $\tilde{\theta} = (D, \beta_0, \beta_1, \alpha r, \gamma, \kappa, \alpha S, C_{\text{init}}, \alpha r_S, C_{\text{pers}})$ donnent la même dynamique de modèle. Changer le coefficient α signifie simplement un changement de l'unité de mesure des rachis colonisés. Nous posons arbitrairement $r_S = 1000$ afin de définir l'unité de mesure du rachis.

Nous avons calculé la distribution a posteriori du paramètre θ par l'algorithme AMIS (*adaptive multiple importance sampling*) développé dans [1]. Cet algorithme est en effet bien adapté lorsque le calcul de la vraisemblance est long (c'est le cas ici car il nécessite la simulation des $(q_a^i(\theta))_{a,i}$ à travers la dynamique de la réaction-diffusion) et pour une distribution cible en forme de banane (une estimation préliminaire du paramètre 2D (β_0, D) montre un tel effet). Cet algorithme a l'avantage de permettre la parallélisation des calculs de vraisemblance des paramètres échantillonnés et de recycler tous les paramètres simulés (ce qui est particulièrement important lorsque la vraisemblance est coûteuse à calculer).

Le principal défi de l'algorithme AMIS est de surmonter la difficulté d'initialisation de l'algorithme. Une étape préliminaire d'estimation grossière des paramètres est effectuée au préalable (à l'aide d'algorithmes de Metropolis-Hasting pour l'estimation de paramètres non fortement corrélés aux autres, et à l'aide de calcul de log-vraisemblances sur une grille grossière). Cette étape préliminaire permet notamment la réduction de modèle suivante:

$$r_S = 1000, \qquad C_{\text{pers}} = 1, \qquad \beta_1 = 0, \qquad r = 0 \tag{12}$$

ainsi qu'une estimation grossière des autres paramètres pour initialiser l'algorithme AMIS.

3 Résultats

L'estimation des paramètres a été réalisée pour plusieurs indices de température. Celui donnant les meilleurs résultats qualitatifs et quantitatifs est l'indice de température où T_a^i est le nombre de jours durant l'été de l'année a où les températures du quadrat ω_i sont au dessus de 28°C. Cet indice permet notamment de décrire le ralentissement de la propagation dans la vallée de la Garonne ainsi que le blocage de l'épidémie dans le sud-est de la France (cf Figure 1 pour la dynamique avec le jeu de paramètre maximisant la vraisemblance).

References

[1] Cornuet, J.-M., Marin, J.-M., Mira, A., and Robert, C. P. (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812.

[2] Grosdidier, M., Ioos, R., Husson, C., Cael, O., Scordia, T., and Marçais, B. (2018). Tracking the invasion: dispersal of *Hymenoscyphus fraxineus* airborne inoculum at different scales. *FEMS Microbiology Ecology*, 94:fiy049.

[3] Grosdidier, M., Ioos, R., and Marçais, B. (2018). Do higher summer temperatures restrict the dissemination of *Hymenoscyphus fraxineus* in France? *Forest Pathology*, 48:e12426.

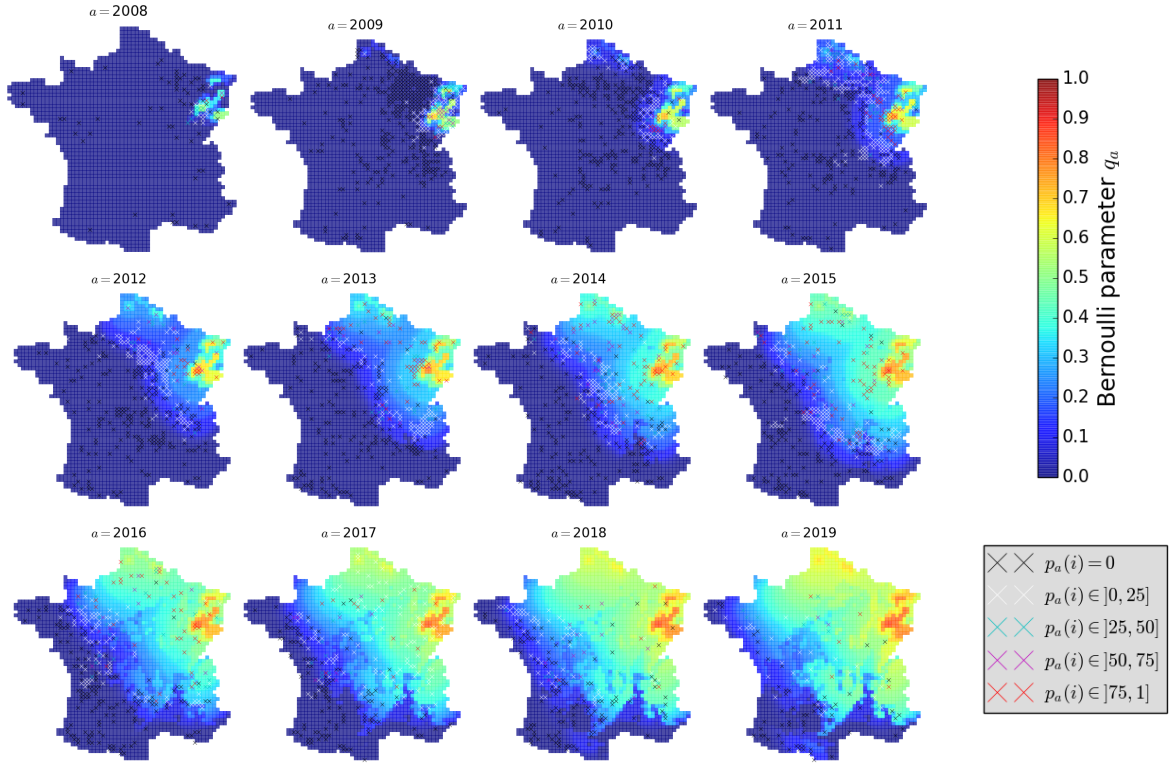


Figure 1: *Heatmap :* Probabilité (8) de développement de symptômes observables pour le paramètre θ qui maximise la vraisemblance. Croix : données d'observations $p_a(i)$ (voir (10)).

[4] Gross, A. and Holdenrieder, O. (2013). On the longevity of *Hymenoscyphus pseudoalbidus* in petioles of *Fraxinus excelsior*. *Forest Pathology*, 43:168–170.

[5] Hauptman, T., Piškur, B., de Groot, M., Ogris, N., Ferlan, M., and Jurc, D. (2013). Temperature effect on *Chalara fraxinea*: heat treatment of saplings as a possible disease control method. *Forest Pathology*, 43:360–370.

[6] Kirisits, T. (2015). Ascocarp formation of *Hymenoscyphus fraxineus* on several-year-old pseudosclerotial leaf rachises of *Fraxinus excelsior*. *Forest Pathology*, 45(3):254–257.

[7] Laubray, S., Buée, M., and Marçais, B. (2023). Evidence of a component Allee effect for an invasive pathogen: *Hymenoscyphus fraxineus*, the ash dieback agent. *Biological Invasions*, 25:2567–2582.

[8] Laubray, S., Buée, M., and Marçais, B. (2024). *Hymenoscyphus fraxineus* persistence in the ash litter. *Plant Pathology*, (in press).

[9] Marçais, B., Giraudel, A., and Husson, C. (2023). Ability of the ash dieback pathogen to reproduce and to induce damage on its host are controlled by different environmental parameters. *PLOS Pathogens*, 19:e1010558.

Modélisation spatio-temporelle de la distribution d'un insecte vecteur

Marie Grosdidier [1], Thomas Opitz [2], Lionel Roques [2] et Emily Walker [2]

[1] *Plateforme ESV, INRAE, BioSP, équipe OPE. France. marie.grosdidier@inrae.fr*
[2] *INRAE, BioSP, équipe recherche. France. emily.walker@inrae.fr*

Résumé. Les modèles de distribution d'espèces sont prépondérants dans les processus de surveillance d'introduction de nouveaux ravageurs ou pathogènes (des plantes). Ils permettent de répondre à des questions de dynamique des populations et peuvent être définis en fonction de l'environnement. Nous cherchons dans cette étude à modéliser la dynamique d'un insecte vecteur, tout en caractérisant son habitat préférentiel. *Monochamus galloprovincialis* est un coléoptère vecteur de *Bursaphelenchus xylophilus*, le nématode du pin. Ce nématode est un organisme de quarantaine prioritaire (OQP) de l'Union européenne à obligation de surveillance officielle annuelle (Règlement UE 2016/2031). La France réalise depuis 2012 une surveillance du vecteur du nématode du pin, *M. galloprovincialis*, par piégeage pour s'assurer du statut exempt du territoire français. La distribution spatio-temporelle de *M. galloprovincialis* en France n'est pas connue et cette information manque crucialement dans les analyses de risque, ainsi que pour la mise en place des stratégies d'échantillonnage. A partir des données de piégeage officielles de 2012 à 2021, la dynamique de population de *M. galloprovincialis* a été estimée spatio-temporellement avec deux approches statistiques : INLA et mécanistico-statistique. L'approche INLA permet l'estimation rapide de modèles hiérarchiques à processus gaussiens latents. L'approche mécanistico-statistique basée sur des chaînes MCMC requiert plus de temps de calcul, mais permet d'estimer le processus d'observation et de la dynamique intra-annuelle. Les connaissances de la biologie de l'insecte vecteur ainsi que les variables environnementales et climatiques d'intérêt ont été intégrées dans ces deux approches complémentaires. Ainsi, les facteurs influençant la population de ce vecteur sont identifiés et peuvent être utilisés pour prédire les zones et les périodes de l'année favorables à la surveillance officielle.

Mots-clés. Vecteur, INLA, approche mécanistico-statistique, estimation bayésienne, statistique spatio-temporelle, modèle de distribution d'espèces

Abstract. Species distribution models are essential to monitor the introduction of new pests or pathogens (of plants). They provide answers to questions about population dynamics and can be defined according to the environment. In this study, we aim modeling the dynamics of an insect vector, while characterizing its preferred habitat. *Monochamus galloprovincialis* is a beetle vector of *Bursaphelenchus xylophilus*, the pinewood nematode. The nematode is a priority quarantine organism (PQO) of the European union with mandatory monitoring annual official (EU Regulation 2016/2031). France has been monitoring the vector of the pinewood nematode, *M. galloprovincialis*, by trapping since 2012 to ensure the free status of the French territory. The spatio-temporal distribution of *M. galloprovincialis* in France is not known and this information is crucially lacking in risk analyses, as well

as for the implementation of sampling strategies. From official trapping data from 2012 to 2021, the population dynamics of *M. galloprovincialis* was estimated spatio-temporally with two statistical approaches: INLA and mechanistic-statistical approach. The INLA modeling allows the rapid estimation of hierarchical models with latent Gaussian processes. The mechanistic-statistical approach based on MCMC chains requires more computing time, but allows to estimate the observation process and the intra-annual dynamics. Knowledge of the biology of the insect vector as well as the environmental and climatic variables of interest were integrated into these two complementary approaches. Thus, factors influencing the population of this vector are identified and can be used to predict areas and times of year favorable for official surveillance.

Keywords. Vector insect, INLA, mecanistico-statistical approach, bayesian estimation, spatio-temporal statistics, species distribution model

1 Introduction

1.1 Situation épidémiologique

L'Union européenne a pour obligation de surveiller le nématode du pin, *Bursaphelenchus xylophilus* (Nickle 1970), organisme de quarantaine prioritaire (OQP, listé dans le règlement UE 2019/1702). Ce nématode provoque la maladie du flétrissement des pins (Pine Wilt Disease, PWD), présente en Europe, au Portugal et en Espagne. La maladie est actuellement absente en France. L'invasion rapide de cette maladie est réalisée par la dissémniation du nématode à travers deux voies de dispersions : une voie de dispersion à longue distance grâce aux matériaux en bois contaminés (Haack, 2006 ; Ye, 2019) et une voie de dispersion à courte distance grâce aux insectes vecteurs du genre Monochamus (Coléoptères). En France, seul *Monochamus galloprovincialis* est capable de vectoriser le nématode du pin (Sousa et al., 2001). La distribution spatio-temporelle de *M. galloprovincialis* est mal connue (Robinet et al., 2024). Ainsi, les analyses de risque et les plans d'échantillonnage pour réaliser la surveillance officielle ne prennent pas en compte cette information pourtant primordiale dans la dynamique de la maladie du flétrissement du pin.

1.2 Modélisation pour l'épidémiologie

Les modèles de distribution d'espèces permettent de répondre à des questions de dynamique des populations et de caractériser l'environnement de ces espèces. Les modèles hiérarchiques, et plus précisèment les modèles à espaces d'état (Durbin et Koopman 2012) permettent de poser un cadre conceptuel où la couche des données est distincte de celle des variables latentes (couche cachée) et de celle des hyper-paramètres. Pour estimer les paramètres et les variables latentes de ce type de modèle hiérarchique, il est souvent plus facile de mettre en oeuvre l'inférence bayésienne. Celle ci permet également de prendre en compte l'incertitude des variables et des paramètres.

1.3 Objectif

L'objectif de ce travail est d'estimer la dynamique de population de *Monochamus galloprovincialis* en France selon sa biologie et l'environnement à partir des données de piégeage de surveillance officielle pour le nématode du pin. Deux approches de modélisation complémentaires sont développées : INLA et mécanistico-statistique.

2 Matériel et Méthode

2.0.1 Variables

Variable à expliquer. Les données de surveillance officielles de *M. galloprovincialis* sont des données d'occurrence (présence et nombre) issus de pièges à phéromones. Au total, 6631 relevés de pièges ont été réalisés sur 10 ans (2012 à 2021), correspondant à 1779 sites de piégeage différents étalés sur l'année entre avril et novembre. Le piège a un rayon d'attractivité de 100 m (Jactel et al., 2019) et est laissé en moyenne 10 jours avant d'être relevé.

Variables explicatives potentielles. Des variables liées aux hôtes (nombre d'arbres, surface, état sanitaire), aux perturbations (incendies, pertes forestières), à l'occupation du sol et aux conditions climatiques (température, précipitations, phénologie des émergences) ont été calculées à différentes échelles spatiales autour des pièges. Le modèle mécanistico-statistique a été paramétré en tenant compte d'une durée de vie moyenne du vecteur de 126 jours (David et al., 2016) et d'une seule génération annuelle (Akbulut et Stamps, 2012).

Sélection des covariables. Une sélection des covariables a été réalisée (random forest, dires d'experts et ACP) pour *in fine* retenir : le taux de mixité dans un rayon de 7500 m autour du piège, les surfaces de pin d'Alep, pin Laricio ou Noir, pin Maritime, pin Sylvestre dans un rayon de 100 m autour du piège, la distance aux feux de forêts de conifères de l'année n-1 le plus proche, la température de la période de piégeage, la somme des précipitations sur la période de piégeage.

2.1 Deux approches de modélisation

2.1.1 Modèle approche INLA

La première approche utilise le package R INLA (integreted nested Laplace Approximation) et permet de tenir compte des dépendances spatiales et temporelles exprimées dans les données tout en estimant les résultats rapidement. Ce premier modèle se base sur une négative binomiale adaptée à une distribution de comptages comportant de nombreux zéros mais aussi beaucoup de fortes valeurs. La variable de réponse Y_{zyt} correspond au nombre de Monochamus capturés par un piège au site $z = 1, \ldots, Z$ l'année $y = 1, \ldots, Y$ à la date de relevé $t = 1, \ldots, T$.

$$Y_{zyt} \sim NegativeBinomial(\mu_{zyt}, E_{zyt})$$

$$\mu_{zyt} = X\beta$$

avec

$$log(\mu_{zyt}) = log(E_{zyt}) + \beta_0 + \sum_{i=1}^{I} \beta_i x_{izyt} + \omega_{zyt}$$

où E_{zyt} est un terme d'exposition connu intégré comme un offset pour pour l'observation zyt(temps de pose du piège), β_0 est l'intercept, et $\beta_i (i = 1, \ldots, M)$ sont les effets fixes reliés aux covariables $x_i (i = 1, \ldots, M)$. Le terme ω_{zyt} fait référence au processus spatiotemporel résiduel captant la variabilité spatiotemporelle non expliquée par les covariables environnementales x_i. Ce processus est structuré avec une dynamique auto-régressive d'ordre 1 et corrélée spatialement. Le Deviance Information Criterion (DIC) a été calculé et est utilisé pour sélectionner le meilleur modèle parmi plusieurs modèles candidats.

2.1.2 Modèle approche mécanistico-statistique

La deuxième approche utilise le package R Nimble avec un modèle défini par une couche latente représentant la dynamique de population modélisée de manière déterministe intra-annuellement, et une couche d'observation modélisée de manière probabiliste/stochastique (Ovaskainen et al., 2008 ; Roques et al., 2011 ; Soubeyrand et Roques 2014). Ce modèle hiérarchique est basé sur un processus de poisson inhomogène. Le nombre de Monochamus piégés Y, pour une période de piégeage t_p pour chaque site y, année z, est modélisé comme suit :

$$Y_{zyt_p} \sim Poisson(\mu_{zyt_p})$$

Le processus d'observation est pris en compte dans l'intensité μ_{zyt}. En effet, l'intensité va dépendre de l'effort d'échantillonnage (temps de piégeage $t_p = t_{releve} - t_{pose}$), de la dynamique d'abondance λ_{zyt}, d'un paramètre de capturabilité du piège ν_0 et du climat (précipitations et températures maximales).

$$\mu_{zyt_p} = \nu_0 * \sum \lambda_{zyt_p} + \nu_1 * log(1 + \sum Precip_{zyt_p}) + \nu_2 * \overline{Tmax}_{zyt_p}$$

La dynamique de population intra-annuelle s'écrit :

$\frac{d\lambda_{zyt}}{dt} = E'(t)$ si $(t - t0_{zy}) < D$, sinon $\frac{d\lambda_{zyt}}{dt} = E'(t - D) - E'(s) * exp^{-\rho * (t - D - s)}$

Cette dynamique dépend :

- d'une fonction d'émergence $E'(t)$ qui prend des valeurs différentes selon le temps t :

% émergence	Interval de temps	Pentes	Valeur $E'(t)$
0-1	avant t0	0	0
1-10	t0-t1	$\alpha 1_{zy}$	$N1_{zy} * \alpha 1_{zy}$
10-90	t1-t2	$\alpha 2_{zy}$	$N1_{zy} * \alpha 2_{zy}$
90-99	t2-t3	$\alpha 3_{zy}$	$N1_{zy} * \alpha 3_{zy}$
99-100	après t3	0	0

avec $t0_{zy}$, $t1_{zy}$, $t2_{zy}$, $t3_{zy}$ les dates d'émergence à 1, 10, 90, 99 % respectivement ; N1 la population disponible ;

- et d'une fonction de mortalité :

$$E'(s) * exp^{-\rho*(t-D-s)}$$

avec ρ le taux de mortalité et D la durée de vie.

$N1$ est calculé comme suit :

$$N1 = N0/(\alpha_1 * (t_1 - t_0) + \alpha_2 * (t_2 - t_1) + \alpha_3 * (t_3 - t_2))$$

où $N0_{zy}$ est la population par jour disponible maximale pour le site z l'année y, exprimée selon : $N0_{zy} = exp(\beta_0 + \sum_{i=1}^{I} \beta_i x_{izy})$, avec β_0 l'intercept et $\beta_i (i = 1, \ldots, M)$ les effets fixes reliés aux covariables $x_i (i = 1, \ldots, M)$.

Le modèle est estimé avec un algorithme bayésien MCMC pour 3 chaînes, 20000 itérations, 1000 burning et 10 thin. Des priors peu informatifs ont été utilisés pour les paramètres du modèles à estimer (β, ν, λ).

3 Résultats et Discussion

Sur les 6631 relevés, 1742 n'ont capturé aucun Monochamus. Concernant les pièges qui ont capturé des Monochamus, une moyenne de 20 Monochamus est observée avec un minimum de 1 et un maximum de 408 Monochamus sur leur période de capture respective. Un pic de capture en juillet est observé avec en plus pour certaines années un second pic en septembre. De plus, l'année 2014 présente un pic de capture en juin. Une différence du nombre de capture entre les pièges positionnés dans la moitié nord de la France et ceux dans la moitié sud est observée avec moitié moins de captures dans la partie nord.

Les deux approches de modélisation permettent de confirmer les informations de la littérature. Le climat joue un rôle primordiale dans la capturabilité des Monochamus par les pièges. La température, la présence d'hôtes et le type d'hôtes (essence) ainsi que les feux de forêts sont des variables qui influent sur la population du vecteur. Nénamoins, avec l'approche INLA une certaine variabilité spatio-temporelle non expliquée par ces covariables est observée et son origine reste encore à découvrir. Par ailleurs, l'approche mécanistico-statistique apporte une information précise sur la dynamique intra-annuelle de la population de Monochamus pour chaque site observé. Grâce à ces deux approches, la dynamique de population de vecteur du nématode du pin en France a pu être expliquée par sa biologie ou par l'environnement. Une projection de ces résultats spatialement et dans les années à venir va permettre de prendre en compte cette population de vecteur dans les analyses de risque et le plan de surveillance du nématode du pin en France ; ces travaux sont réalisés dans le cadre du GT surveillance du nématode du pin de la Plateforme ESV.

Bibliographie

Akbulut, W.T. Stamps (2011). Insect vectors of the pinewood nematode: a review of the biology and ecology of Monochamus species. *For. Pathol.*, 42, pp. 89-99.

David, G., Giffard, B., Piou, D., Roques, A., Jactel, H. (2016). Potential effects of climate warming on the survivorship of adult *Monochamus galloprovincialis*, *Agricultural & Forest Entomology*, 19, pp. 192-199.

Durbin, J., Koopman, SJ. (2012). Time series analysis by state space methods, vol 38. *Oxford University Press*, Oxford.

Haack, R.A. (2006). Exotic bark- and wood-boring Coleoptera in the United States: recent establishments and interceptions, *Revue Canadienne De Recherche Forestière*, 36, pp. 269-288.

Jactel, H., Bonifacio, L., van Halder, I., Vétillard, F., Robinet, C. and David, G. (2019). A novel, easy method for estimating pheromone trap attraction range: application to the pine sawyer beetle *Monochamus galloprovincialis*. *Agr Forest Entomol*, 21, pp. 8-14.

Koutroumpa, F. A. (2008). Biologie et phylogeographie de *Monochamus galloprovincialis* (coleoptera, cerambycidae) vecteur du nématode du pin en Europe. Thèse. Interactions entre organismes. Université d'Orléans.

Lee, T., Kim, J. (2025). Rapid spread and high prevalence of the Pine Wilt Disease around wildfire areas. *Trees, Forests and People*. Available at SSRN: https://ssrn.com/abstract=5087606.

Ovaskainen O, Rekola H, Meyke E, Arjas E (2008). Bayesian methods for analyzing movements in heterogeneous landscapes from mark-recapture data. *Ecology*, 89(2), pp. 542–554.

Naves, P., de Sousa, E. (2009). Threshold temperatures and degree-day estimates for development of post-dormancy larvae of *Monochamus galloprovincialis* (Coleoptera: Cerambycidae). *J Pest Sci*, 82, pp. 1–6.

Robinet, C., Raffin, A., Jactel, H., Kersaudy, E., Deuffic, P., Clopeau, A., Hotte, H., Kleinhentz, M., Robin, C., Roux, G., Villate, L., Castagnone-Sereno, P. (2024). Improving monitoring and management methods is of the utmost importance in countries at risk of invasion by the pinewood nematode. *Annals of Forest Science*, 81, pp. 16.

Roques, L., Soubeyrand, S., Rousselet, J. (2011). A statistical-reaction–diffusion approach for analyzing expansion processes. *Journal of Theoretical Biology*, 274, pp. 43–51.

Soubeyrand, S., Roques, L. (2014). Parameter estimation for reaction-diffusion models of biological invasions. *Population ecology*, 56, pp. 427-434.

Sousa, E., Bravo, M.A., Pires, J., Naves, P., Penas, A.C., Bonifacio, L., Mota, L. (2001). Bursaphelenchus xylophilus (Nematoda; Aphelenchoididae) associated with *Monochamus galloprovincialis* (Coleoptera; Cerambycidae) in Portugal, *Nematology*, 3, pp. 89-91.

Ye, J.R. (2019). Episemic status of pine wilt disease in China and its prevention and control techniques and counter measures, *Scientia Silvae Sinicae*, 55, pp. 1-10.

Arbocarto, outil de modelisation de la dynamique des populations des moustiques *Aedes* : un vecteur de dialogue entre recherche, développement, et lutte antivectorielle

Renaud Marti [1] Pachka Hammami [2] & Mathieu Castets [1] & Marie Demarchi[4] & Grégory L'Ambert[5] & Guillaume Lacour[6] & Thierry Baldet[2] & Annelise Tran [1,2]

[1] *TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France, renaud.marti@inrae.fr, annelise.tran@cirad.fr*
[2] *ASTRE, Univ Montpellier, CIRAD, INRAE, Montpellier/Ste-Clotilde, France, pachka.hammami@cirad.fr*
[4] *Marie Demarchi, Maison de la Télédétection, Montpellier, France*
[5] *EID Méditerranée, Montpellier, France*
[6] *ALTOPICTUS, Bayonne/Pérols, France*

Résumé. A la demande de la Direction Générale de la Santé (DGS), un outil générique dénommé ARBOCARTO a été développé, afin de simuler la dynamique de populations de moustique du genre *Aedes* en milieu tropical et tempéré. Son développement s'est appuyé sur un long historique de collaboration entre les instituts de recherche (CIRAD principalement) et les acteurs de la lutte antivectorielle (EID, ARS). La poursuite des travaux académiques permet d'intégrer régulièrement de nouvelles fonctionnalités dans l'outil, toujours en dialogue avec les services opérationnels impliqués.

Mots-clés. Modélisation mécaniste, dynamique de population, moustiques vecteur

Abstract. At the request of the Direction Générale de la Santé (DGS), a generic tool called ARBOCARTO has been implemented to simulate the population dynamics of *Aedes* mosquitoes in both tropical and temperate environments. Its development was based on a long history of collaboration between research institutes (mainly CIRAD) and vector control stakeholders (EID, ARS). The continuation of academic work means that new functionalities are regularly integrated into the tool, always in dialogue with the operational services involved.

Keywords. Process-based modelling, poulation dynamic, mosquitoes, vector mosquito

1 Un modèle généraliste co-construit

1.1 Deux espèces invasives, antropophiles et vectrices d'arboviroses

Les moustiques *Aedes albopictus* "moustique tigre" et *Aedes aegypti* sont les principaux vecteurs des virus de la dengue, du chikungunya, du Zika, et sont deux espèces pleinement

adaptées aux milieux urbains. Leur haute plasticité écologique leur a permis de coloniser de nombreuses régions du monde en tirant bénéfice d'une mondialisation des échanges, des processus d'urbanisation des terres, et des grands changements environnementaux et climatiques en cours (Kraemer et al, 2015). Sur le territoire français, *Ae. albopictus* est présent dans l'Océan Indien (Réunion et Mayotte) et dans 81 départements de France hexagonale (2024). *Aedes aegypti* est présent aux Antilles, en Guyane et dans l'Océan indien. La présence de ces espèces est à l'origine d'épidémies dans les différents territoires outre-mer et constitue un risque important également pour la France hexagonale, où le moustique "tigre" est implanté depuis 2004 (Cochet et al., 2022).

Dans ce contexte sanitaire préoccupant, la Direction Générale de la Santé (DGS) a sollicité les instituts de recherche, principalement le CIRAD, pour développer un outil de modélisation des dynamiques de populations de moustique "tigre" applicable à différents contextes géographiques du territoire français.

1.2 Un modèle démographique simple fondé sur le cycle de vie du moustique

La modélisation mise en place est fondée sur une approche mécaniste, consistant à décrire les principaux états physiologiques de l'espèce de moustique étudiée, et à documenter les facteurs et les processus conduisant le passage d'un état à un autre. Tous les moustiques ont en commun un cycle de vie qui commence par un stade aquatique immature avec trois grandes étapes de développement : œuf, larve, puis nymphe (Fig. 1). Le bilan démographique propre à chacun des stades peut être décrit par un système d'équations différentielles ordinaires (EDO), qui permet de quantifier chacun des 10 états populationnels au cours du temps : 3 états aquatiques (E œufs, L larves, P pupes), 1 état émergent (Aem), 3 états nullipares (A1h, A1g, A1o) et 3 états pares (A2h, A2g, A2o). La dynamique temporelle est guidée par les variables météorologiques (température et précipitation) dont dépendent les fonctions de transitions entre les stades physiologiques (Marti et al 2022).

Les variations spatiales d'abondance des populations de moustiques sont contraintes par la distribution des ressources, en particulier par la densité de gîtes potentiels pour la ponte et le développement des stades larvaires. Ces gîtes sont de faibles dimensions pour *Ae. aegypti* et *Ae. albopictus* (coupelles, pneus, cavités, bidons fûts, etc.). Dans l'approche ARBOCARTO, l'espace de surveillance et de contrôle entomologique est découpé au préalable en unités géographiques distinctes, de quelques hectares, considérées comme indépendantes (métapopulations de type "îles"). Une valeur de "capacité de charge de l'environnement", ici définie comme le nombre maximal théorique de larves, est estimée pour chaque unité. Cette estimation s'appuie sur une connaissance experte du terrain transmise par les services de lutte antivectorielle. A défaut, elle peut être inférée par une vision de géomaticien/télédétetecteur, en associant des classes d'occupation et d'usage du sol à des valeurs pré-établies de capacité de charge larvaires, par comparaison à d'autres terrain d'étude mieux documentés, comme La Réunion (Hoarau 2020, Teillet 2024).

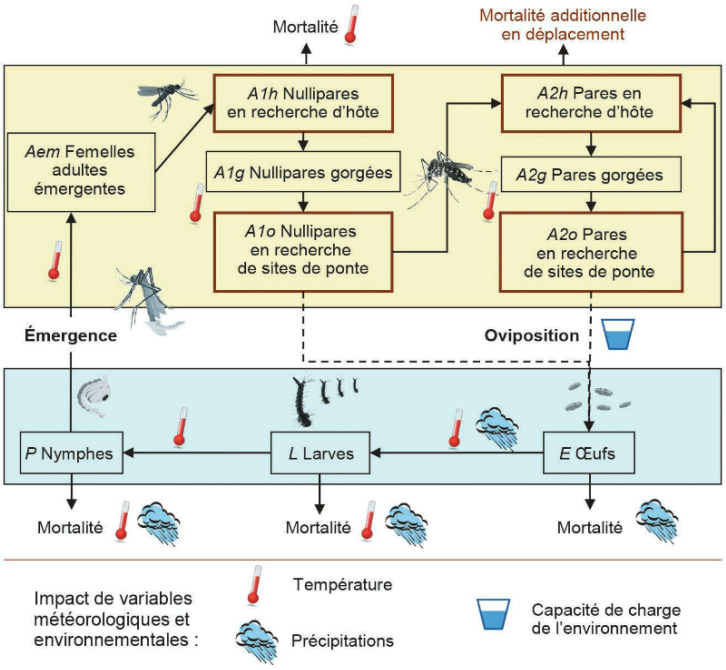


Figure 1: Cycle de vie du moustique représenté sous la forme de compartiments.

1.3 Un dialogue dans la durée entre chercheurs et opérateurs de démoustication

La définition des états et des processus retenus est le fruit d'un échange entre entomologistes, acteurs de la lutte antivectorielle, et modélisateurs. Leur paramétrisation s'appuie sur la littérature et/ou la connaissance d'experts. Si certains mécanismes ont été simplifiés (1 seul stade larvaire au lieu de 4), ou ignorés en l'état actuel du modèle (déplacement des moustiques entre unités géographiques), le modèle distingue l'abondance de moustiques (femelles) adultes susceptibles de prendre un repas sanguin (A1h et A2h), et de transmettre un agent pathogène après une première piqûre infectante (A2h).

Les premières implémentations de ce modèle générique ont été effectuées en Camargue pour des espèces appartenant aux 3 genres *Anopheles*, *Culex*, et *Aedes* (Cailly et al 2012 ; Ezanno et al 2015). La poursuite de cette collaboration entre le CIRAD et l'EID Méditerranée a permis d'établir un jeu de paramètres et de fonctions de transition spécifiques pour l'espèce *Ae. albopictus* en milieu tempéré, calibré et validé sur la ville de Nice (Tran et al, 2013), tandis qu'une collaboration similaire avec l'ARS Réunion a mené à la co-construction de l'application AlboRun (Tran et al. 2020) applicable en milieu tropical. A la demande de la Direction Générale de la Santé (DGS), l'outil ARBOCARTO a été créé, afin d'être utilisable dans les différents contextes (Fig. 2 3). Les actions de formation sont régulièrement assurées (France hexagonale, départements et régions insulaires d'outre-mers de La Réunion, Guadeloupe, Martinique, Mayotte, région Guyane, ou encore plus récemment Brésil), ce qui permet de bénéficier de retours utilisateurs et de faire évoluer l'outil.

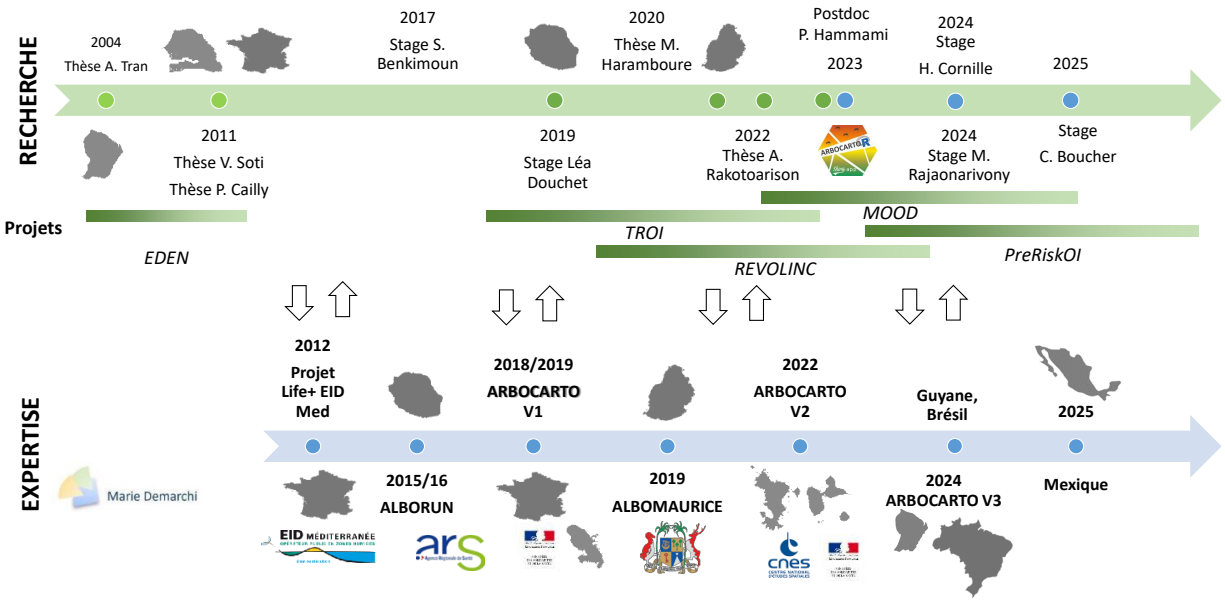


Figure 2: Chronologies des échanges entre le développement d'ARBOCARTO (*Ae. albopictus et aegypti*) au travers de travaux de recherche et ses déclinaisons opérationnelles selon les régions.

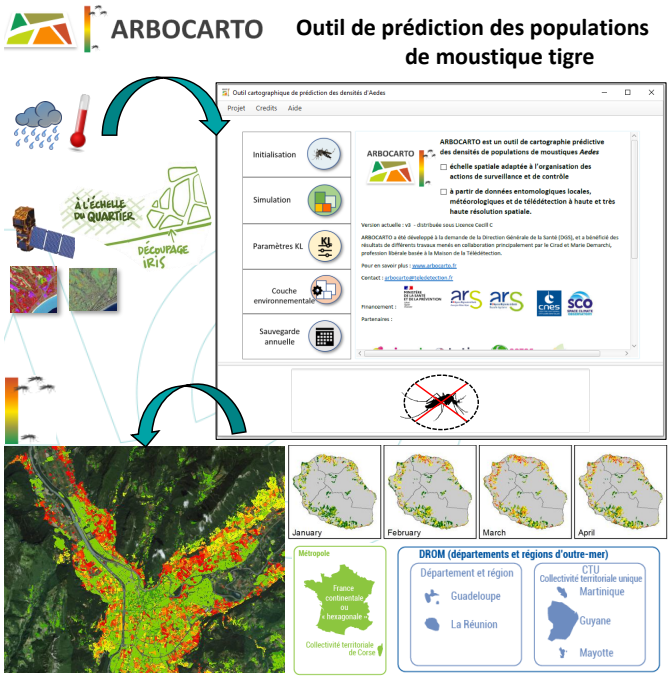


Figure 3: Interface, entrées et sorties de l'outil ARBOCARTO.

2 Evolutions du modèle vers un "écosystème" ARBOCARTO

Sur le plan académique, les travaux effectués lors de stages, contrats doctoraux et postdoctoraux permettent de développer régulièrement de nouvelles fonctionnalités qui sont peu à peu intégrées à l'application, comme le calcul du taux de reproduction de base du virus de la dengue selon la formule de Ross-Macdonald (R0, Fig. 2 : Benkimoun 2017), ou l'intégration de la technique de l'insecte stérile (éventuellement boostée) aux méthodes de lutte antivectorielle initialement proposées dans ARBOCARTO (destruction de gîtes, biocides) (Fig. 2 : Douchet 2019 ; Haramboure 2020).

Le couplage de la sortie du modèle déterministe ARBOCARTO avec un modèle stochastique à compartiments, SEIR (hôtes) et SEI (vecteurs), pour la dengue, Zika, et chikungunya, a permis de construire une application en ligne directement accessible aux utilisateurs : (Fig. 2 et 4 : Hammami 2023). Ce modèle, désigné comme arbocartoR, prend également en compte la mobilité quotidienne des hôtes (navettage). Pour évaluer les futurs effets du changement climatique sur *Aedes albopictus* à La Réunion, le modèle ARBOCARTO a été couplé en entrée avec les sorties climatiques du projet BRIO, ce qui permet de projeter la dynamique de population du moustique selon différents scénarios socio-économiques à horizon 2100 à une haute résolution spatiale (3 km) (Lamy et al. 2023 ; Fig. 2 et 4 : Rajaonarivony 2024).

Si ARBOCARTO est actuellement centrée sur deux espèces vectrices du genre *Aedes*, l'approche générique de modélisation permet d'envisager d'intégrer dans l'interface des espèces de moustiques d'intérêt médical déjà documentés (ou encore à modéliser) appartenant aux genres *Anopheles*, vecteurs du paludisme (Rakotoarison, 2022), ou *Culex*, vecteur de la fièvre du Nil Occidental ou de la fièvre de la Vallée du Rift (Hammami et al 2016 ; Tran et al 2019).

Parmi les hypothèses simplificatrices du modèle, la notion de gîtes de repos n'apparaît pas explicitement. Un gîte de repos peut être défini comme un abri permettant une régulation thermique, une protection contre les prédateurs ou une exposition directe au soleil, notamment pour les femelles qui sont à la croisée des états gorgés et gravides (Ag), et qui doivent faire maturer leurs oeufs. Les outils de la géoinformatique permettent de calculer à partir de données LIDAR un indice de densité folaire (*Leaf area index*), que l'on peut associer à un effet microclimatique de la végétation. Cet indice pourrait être potentiellement représentatif de la qualité des gîtes de repos végétalisés pour les moustiques, et pourrait être intégré à terme dans l'application ARBOCARTO (4).

La prise en compte du déplacement du moustique entre unités géographiques proches présente un intérêt pour lancer des simulations sur des parcelles de très petite taille et/ou dans un contexte géographique peu favorable (zones d'altitude), ainsi que pour simuler les phénomènes de recolonisation après une forte action de LAV. Cet aspect sera bientôt implémenté pour construire des scénarios d'invasion de La Réunion par *Anopheles stephensi* en considérant ce phénomène de dispersion active, tout en considérant également la dispersion passive (p.ex. réseau routier) (Fig. 3 : Bouchet 2025).

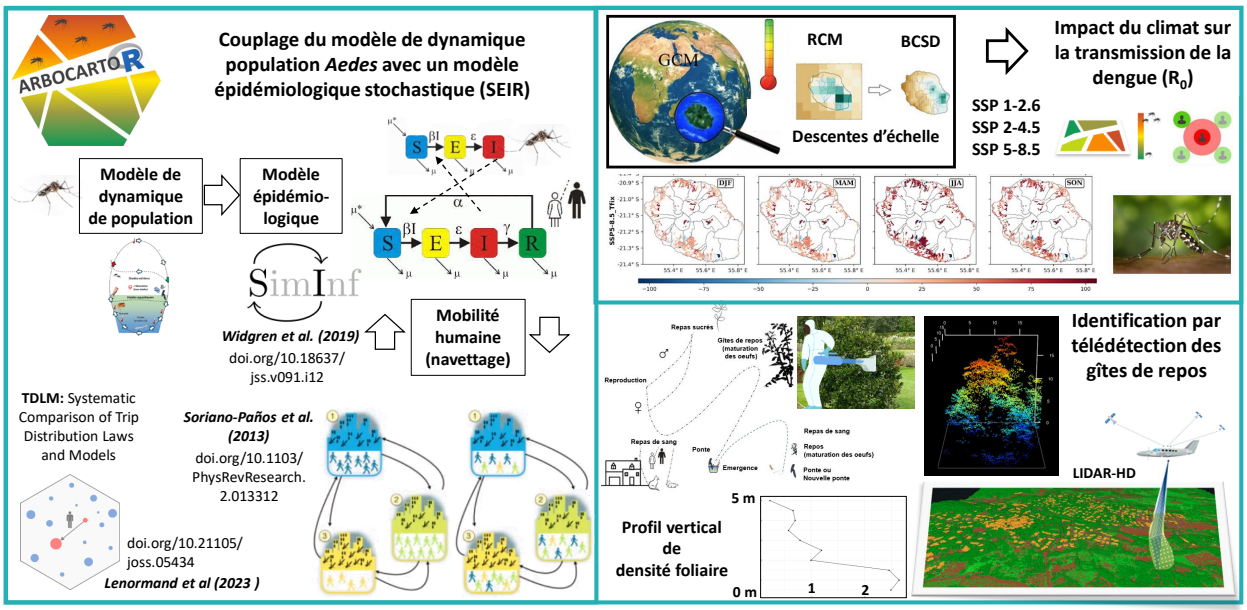


Figure 4: Travaux en cours, "ecosystème" Arbocarto.

Conclusion

Le développement d'ARBOCARTO a bénéficié d'un important historique de collaboration entre instituts de recherche (CIRAD) et acteurs impliqués dans la surveillance et le contrôle des moustiques vecteurs (en particulier, *Ae. albopictus* et *Ae. aegypti*). Il s'est également appuyé sur un certain nombre d'hypothèses de modélisation simplificatrices guidées par l'état des connaissances, et un pragmatisme lié au besoin d'opérationnalité. Les travaux en cours permettent de maintenir actif le dialogue entre recherche et acteurs de la lutte antivectorielle.

Bibliographie

Cailly, P., Tran, A., Balenghien, T., L'Ambert, G., Toty, C., & Ezanno, P. (2012). A climate-driven abundance model to assess mosquito control strategies. Ecological Modelling, 227, 7-17.

Cochet, A., Calba, C., Jourdain, F., Grard, G., Durand, G. A., Guinard, A., ... & Franke, F. (2022). Autochthonous dengue in mainland France, 2022: geographical extension and incidence increase. Eurosurveillance, 27(44), 2200818.

Ezanno, P., Aubry-Kientz, M., Arnoux, S., Cailly, P., L'Ambert, G., Toty, C., ... & Tran, A. (2015). A generic weather-driven model to predict mosquito population dynamics applied to species of Anopheles, Culex and Aedes genera of southern France. Preventive veterinary medicine, 120(1),

39-50.

Hammami P., Annelise, T., Alan, K. et al. Rift Valley fever vector diversity and impact of meteorological and environmental factors on Culex pipiens dynamics in the Okavango Delta, Botswana. Parasites Vectors 9, 434 (2016). https://doi.org/10.1186/s13071-016-1712-1

Hoarau, O. (2021). Télédétection et modélisation de populations de moustiques vecteurs: Extraction d'indicateurs paysagers à partir d'images d'Observation de la Terre pour l'estimation de la distribution des gîtes larvaires en milieu urbain (Doctoral dissertation, Université de la Réunion).

Lamy, K., Tran, A., Portafaix, T., Leroux, M. D., & Baldet, T. (2023). Impact of regional climate change on the mosquito vector Aedes albopictus in a tropical island environment: La Réunion. Science of the Total Environment, 875, 162484.

Marti, R., Demarchi, M., Castets, M., & Tran, A. (2022). Chapitre 8 - *Arbocarto : un modèle mécaniste fondé sur le cycle de vie des moustiques Aedes.* In A. Tran, E. Daudé, & T. Catry (éds.), Télédétection et modélisation spatiale. Quæ https://books.openedition.org/quae/42909

Rakotoarison, H. A. (2022). Télédétection et modélisation spatiale: application à la surveillance et au contrôle du paludisme à Madagascar (Doctoral dissertation, Université de Montpellier).

Teillet, C., Devillers, R., Tran, A., Catry, T., Marti, R., Dessay, N., ...& Roux, E. (2024a). Exploring fine-scale urban landscapes using satellite data to predict the distribution of Aedes mosquito breeding sites. International Journal of Health Geographics, 23(1), 18. https://doi.org/10.1186/s12942-024-00378-3

Tran, A., L'ambert, G., Lacour, G., Benoît, R., Demarchi, M., Cros, M., ... & Ezanno, P. (2013). A rainfall-and temperature-driven abundance model for Aedes albopictus populations. International journal of environmental research and public health, 10(5), 1698-1719.

Tran, A., Fall, A. G., Biteye, B., Ciss, M., Gimonneau, G., Castets, M., ... & Chevalier, V. (2019). Spatial modeling of mosquito vectors for Rift Valley Fever virus in Northern Senegal: integrating satellite-derived meteorological estimates in population dynamics models. Remote Sensing, 11(9), 1024.

Tran, A., Mangeas, M., Demarchi, M., Roux, E., Degenne, P., Haramboure, M., ... & Dehecq, J. S. (2020). Complementarity of empirical and process-based approaches to modelling mosquito population dynamics with Aedes albopictus as an example—Application to the development of an operational mapping tool of vector populations. PloS one, 15(1), e0227407.

Modèles de mélange

Semi-parametric estimation of location-scale mixture densities

Mohamed El Hasnaoui [1] & Matthieu Marbac [2]

[1] *Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France, mohamed.elhasnaoui@ensai.fr*

[2] *Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France, matthieu.marbac-lourdelle@ensai.fr*

Résumé. Nous proposons un estimateur semi-paramétrique pour les densités de mélange. La densité estimée appartient à une famille position-échelle où les composantes diffèrent uniquement par des facteurs de translation et d'échelle. En utilisant la M-estimation, nous estimons les proportions du mélange, les paramètres de translation et d'échelle, ainsi que la densité des composantes. La consistance découle de l'identifiabilité et de la compacité de l'espace des paramètres, tandis que le taux de convergence est obtenu en approximant la densité commune dans un sous-espace de dimension finie de $L^2(\mathbb{R})$ dont la dimension croît avec la taille de l'échantillon. Nous calculons l'estimateur en utilisant un algorithme d'Expectation-Maximization généralisée, et illustrons les résultats avec des simulations numériques.

Mots-clés. Processus empirique, Modèle de mélange, Sélection de modèle

Abstract. We propose a semi-parametric estimator for mixture densities. The estimated density belongs to a location-scale mixture family where component densities differ only by translation and scaling factors. Using M-estimation, we estimate the mixture proportions, location and scale parameters, and the component density. Consistency follows from parameter identifiability and parameter space compactness, while the convergence rate follows from approximating the common component density in a finite-dimensional subspace of $L^2(\mathbb{R})$ whose dimension increases with the sample size. We derive the estimator using a Generalized Expectation-Maximization (GEM) algorithm, then illustrate our results with numerical simulations.

Keywords. Empirical process, Location-scale mixture model, Model selection

1 Introduction

Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed real-valued random variables following a probability distribution characterized by an unknown density $f^\star$ with respect to Lebesgue's measure. We assume that $f^\star$ is an element of a location-scale mixture family of distributions with K components. More precisely, $f^\star$ takes the form $f^\star = f_{K,\boldsymbol{\theta}_K^\star}$ such that

$$f_{K,\boldsymbol{\theta}}(x) = \sum_{k=1}^{K} \frac{\pi_k}{\sigma_k} \psi\left(\frac{x-\mu_k}{\sigma_k}\right),$$

where $\boldsymbol{\theta} = (\boldsymbol{\upsilon}, \psi)$ groups the finite-dimensional parameters $\boldsymbol{\upsilon} = (\boldsymbol{\pi}^T, \boldsymbol{\mu}^T, \boldsymbol{\sigma}^T)^T \in \Upsilon_K \subset \mathbb{R}^{3K}$ and the symmetric probability density function $\psi \in \mathcal{H} \subset L^2(\mathbb{R})$, where $L^2(\mathbb{R})$ the set of square-integrable functions over $\mathbb{R}$. The parameter space is $\Theta_K = \Upsilon_K \times \mathcal{H}$.

In the literature, the various contributions are mainly concerned with the case $K = 2$ or 3 without scale parameters. Showing identifiability for $K = 2$ or 3, Hunter et al. (2004) use a non-parametric approach, where the estimate of the vector $(\boldsymbol{\pi}^T, \boldsymbol{\mu}^T)$ is the minimizer of an L_p-distance and the estimate of ψ is based on the empirical cumulative distribution function of the sample. Butucea et al. (2014) treat the problem using M-estimation based on a Fourier approach.

We generalize these results in the following way. We propose estimating ψ by estimating its coefficients on a basis of $L^2(\mathbb{R})$. By working on a finite number of elements B of the basis, we reduce our semi-parametric problem into a parametric one. Assuming the identifiability of the true parameter, we obtain an M-estimator of the vector of finite-dimensional parameters along with the basis coefficients. This estimator is consistent for a fixed B. Its rate of convergence follows after determining the rates of its variance and bias. To control the variance, our argument is based on a Taylor expansion and the dependence of the Berry-Esseen bound on the dimension B (see Bentkus (2003)). This variance converges to zero when B tends to infinity in an appropriate rate. To control the bias, additional regularity conditions, as in Belomestny et al. (2019), allow us to conclude. We use the Generalized Expectation-Maximization algorithm (GEM algorithm) to compute our estimator and illustrate the theoretical findings with numerical experiments.

In Section 2, we introduce our theoretical estimation framework. We introduce the parameter space and some of its properties along with the loss function. Then, we present our consistency results. In Section 3, we describe the practical aspects of the estimation concerning the GEM algorithm. In Section 4, we illustrate our results on several examples. Finally, in Section 5, we conclude and discuss the ongoing work on convergence rates.

2 Location-scale mixtures estimation in theory

2.1 Statistical framework

M-estimation involves minimizing a loss. To guarantee the existence of the various minimizers, we define Υ_K and $\mathcal{H}$ to be compact sets, so that the parameter space Θ_K is compact.

First, we ensure that $\boldsymbol{\pi}$ represents a probability mass function, it belongs to the K-dimensional simplex $\mathcal{S}_K$. Second, $\boldsymbol{\mu}$ belongs to a hypercube; there exists $M > 0$ such that every translation parameter μ_k belong to $[-M, M]$. Third, $\boldsymbol{\sigma}$ belongs to another hypercube, guaranteeing the scales are positive and bounded; there exists $s, S > 0$ such that every scale parameter σ_k belongs to $[s, S]$. Thus, we define $\Upsilon_K = \mathcal{S}_K \times [-M, M]^K \times [s, S]^K$.

Next, we take $\mathcal{H}$ as a subset of $L^2(\mathbb{R})$ to bypass the infinite-dimensional parameter estimation hurdle. Using the Arzelà-Ascoli theorem, we ensure the compactness of the functional space $\mathcal{H}$ by defining it as a uniformly bounded and uniformly equicontinuous family of func-

tions. We choose $\mathcal{H}$ as a family of Lipschitz-continuous functions, with a common Lipschitz constant $L > 0$. To control the variations of the loss we minimize for the consistency, we introduce two functions $u : \mathbb{R} \mapsto \mathbb{R}$, continuous and integrable, and $l : \mathbb{R} \mapsto \mathbb{R}$, positive, to uniformly bound all $\psi \in \mathcal{H}$. We assume that u and l verify the integrability properties $\int_{-\infty}^{\infty} u(x)|\log u(x)|dx < \infty$ and $\int_{-\infty}^{\infty} u(x)|\log l(x)|dx < \infty$. These functions not only bound the elements ψ of the space, but also bound each of their location-scale transformations $\psi(\frac{-\mu}{\sigma})$ such that $(\mu, \sigma) \in \Upsilon = [-M, M] \times [s, S]$. For brevity, let $u_{inf}(x) = \inf_{(\mu,\sigma)\in\Upsilon} u(\sigma x + \mu)$ and $l_{sup}(x) = \sup_{(\mu,\sigma)\in\Upsilon} l(\sigma x + \mu)$. Finally, we also suppose $u(x) \to 0$ as $|x| \to \infty$, which does not restrict the space $\mathcal{H}$, as all probability density functions tend to zero at infinity. We define $\mathcal{H}$ as the following set of real-valued functions over $\mathbb{R}$:

$$\mathcal{H} = \left\{\psi \in \mathcal{F} \cap \mathcal{C}(\mathbb{R}, L) \cap L^2(\mathbb{R}) : \quad l_{sup} \leq \psi \leq u_{inf}\right\},$$

where $\mathcal{F}$ the set of symmetric densities over $\mathbb{R}$ and $\mathcal{C}(\mathbb{R}, L)$ the set of L-Lipschitz continuous functions over $\mathbb{R}$.

We define the parameter space by $\Theta_K = \Upsilon_K \times \mathcal{H}$ and view it as a metric space with the distance function $(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \mapsto \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_\infty$, where $\|.\|_\infty$ is defined for $\boldsymbol{\theta} = (\boldsymbol{\upsilon}, \psi) \in \Theta_K$ by $\|\boldsymbol{\theta}\|_\infty = \max(\|\boldsymbol{\upsilon}\|_\infty, \|\psi\|_\infty)$, where $\|\boldsymbol{\upsilon}\|_\infty = \max_{i=1,\cdots,3K} |\upsilon_i|$ and $\|\psi\|_\infty = \sup_{x\in\mathbb{R}} |\psi(x)|$.

Our goal is to find $\boldsymbol{\theta} \in \Theta_K$ that minimizes the Kullback-Leibler divergence with respect to the true distribution, or equivalently, the following loss function

$$\mathcal{L}(\boldsymbol{\theta}) = -\int_{-\infty}^{\infty} \log f_{K,\boldsymbol{\theta}}(x) f^\star(x) dx.$$

To manage the "label-switching problem" (Section 2, Hunter et al. (2004)), which is linked to identifiability, we define the function $\mathcal{P}_K : \Upsilon_K \mapsto \Upsilon_K$ that associates to each $\boldsymbol{\upsilon}$ a vector $\boldsymbol{\upsilon}_\rho$, where ρ is the permutation that orders the scaling parameters from the largest to the smallest, and orders the location parameters in a similar way for equal scaling parameters (such permutation depends on $\boldsymbol{\upsilon}$). With a slight abuse of notation, we denote by $\mathcal{P}_K(\boldsymbol{\theta})$ the vector $(\mathcal{P}_K(\boldsymbol{\upsilon}), \psi)$. Applying the function $\mathcal{P}_K$ does not alter the value of the loss function: for any $\boldsymbol{\theta} \in \Theta_K$, $\mathcal{L}(\mathcal{P}_K(\boldsymbol{\theta})) = \mathcal{L}(\boldsymbol{\theta})$. We will assume that the true parameter $\boldsymbol{\theta}_K^\star = (\boldsymbol{\pi}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\sigma}^\star, \psi^\star)$ is an element of Θ_K and that, without loss of generality, it satisfies $\pi_1^\star, \cdots, \pi_K^\star > 0$, $\sigma_k^\star > \sigma_{k+1}^\star$ and $\mu_k^\star > \mu_{k+1}^\star$ when $\sigma_k^\star = \sigma_{k+1}^\star$, for every $k \in \{1, \cdots, K-1\}$. In particular, $\mathcal{P}_K(\boldsymbol{\theta}_K^\star) = \boldsymbol{\theta}_K^\star$.

2.2 Consistency of the M-estimator

Given a finite sample $X_1, \cdots, X_n$, we minimize $\mathcal{L}_n$, the empirical version of $\mathcal{L}$, defined by

$$\mathcal{L}_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \log f_{K,\boldsymbol{\theta}}(X_i).$$

Any minimizer $\widehat{\boldsymbol{\theta}}_{K,n}$ of $\mathcal{L}_n$ over Θ_K is a consistent estimator of $\boldsymbol{\theta}_K^\star$ in the following sense.

Theorem 1. *Assuming the true parameter $\boldsymbol{\theta}_K^\star$ is identifiable: if $\boldsymbol{\theta} \in \Theta_K$ and $f_{K,\boldsymbol{\theta}} = f_{K,\boldsymbol{\theta}_K^\star}$ almost everywhere, then $\mathcal{P}_K(\boldsymbol{\theta}) = \boldsymbol{\theta}_K^\star$, any minimizer $\widehat{\boldsymbol{\theta}}_{K,n}$ of $\mathcal{L}_n$ over Θ_K is a consistent estimator of $\boldsymbol{\theta}_K^\star$ under the norm $\|.\|_\infty$:*

$$\|\mathcal{P}_K(\widehat{\boldsymbol{\theta}}_{K,n}) - \boldsymbol{\theta}_K^\star\|_\infty = o_{\mathbb{P}}(1).$$

There is one downside for the estimator $\widehat{\boldsymbol{\theta}}_{K,n}$; we cannot determine its rate of convergence unless we restrain the parameter space further, as it is a subset of an infinite-dimensional vector space. Our focus will be on an estimator for the common component density defined in a truncated functional space $\mathcal{H}_B \subset \mathcal{H}$.

We let $(\omega_b)_{b\in\mathbb{N}}$ be an orthonormal basis for $L^2(\mathbb{R})$. Since $\mathcal{H} \subset \mathrm{L}^2(\mathbb{R})$, for each $\psi \in \mathcal{H}$, there is a unique sequence of real numbers $(\beta_b)_{b\in\mathbb{N}}$ in $\ell^2(\mathbb{N})$, the space of square-summable real-valued sequences, such that, for almost all $x \in \mathbb{R}$, one has $\psi(x) = \sum_{b=0}^{\infty} \beta_b \omega_b(x)$. The Hermite function basis is an example of such basis, where we have

$$\omega_b(x) = \frac{H_b(x)e^{-\frac{x^2}{2}}}{\sqrt{2^b b!\sqrt{\pi}}},$$

with H_b the Hermite polynomial of order b, defined by $H_b(x) = (-1)^b e^{x^2} \frac{d^b}{dx^b} e^{-x^2}$.

For $B \in \mathbb{N}$, we refer to the decomposition above to define the subset $\mathcal{H}_B$ as $\mathcal{H}_B = \{\psi \in \mathcal{H} : \forall b > B, \quad \beta_b = 0\}$. The corresponding truncated parameter space is $\Theta_{K,B} = \Upsilon_K \times \mathcal{H}_B$ and we have $\Theta_{K,B} \subset \Theta_{K,B+1} \subset \Theta_K$.

The following consistency result is valid in the truncated parameter space.

Theorem 2. *For a fixed $B \in \mathbb{N}$, under mild assumptions on the basis functions guaranteeing the generic identifiability of the minimizer $\boldsymbol{\theta}^{\star}_{K,B}$ of $\mathcal{L}$ over $\Theta_{K,B}$, any minimizer $\widehat{\boldsymbol{\theta}}_{K,B,n}$ of $\mathcal{L}_n$ over $\Theta_{K,B}$ is a consistent estimator of $\boldsymbol{\theta}^{\star}_{K,B}$ in the following sense*

$$\|\mathcal{P}_K(\widehat{\boldsymbol{\theta}}_{K,B,n}) - \boldsymbol{\theta}^{\star}_{K,B}\|_{\infty} = o_{\mathbb{P}}(1).$$

3 Inference via the Generalized EM Algorithm

We discussed estimation for a fixed B. In practice, we recommend using the Bayesian Information Criterion (BIC) to determine an appropriate value of $B \in \{1, \cdots, B_{max}\}$, where B_{max} is a fixed positive integer. B_{max} depends on the sample size n, and this dependence is dictated by the convergence rate result.

As previously mentioned, we estimate the parameters by minimizing the negative log-likelihood function $\mathcal{L}_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \frac{\pi_k}{\sigma_k}\psi\left(\frac{X_i-\mu_k}{\sigma_k}\right)\right)$. We opt for a Generalized Expectation-Maximization algorithm (GEM algorithm) to minimize the negative complete-data likelihood expressed as follows

$$\mathcal{L}_n^c(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} Z_{i,k} \log\left(\frac{\pi_k}{\sigma_k}\psi\left(\frac{X_i-\mu_k}{\sigma_k}\right)\right),$$

where the values of $Z_{i,k}$ are the missing data (See Dempster et al. (1977) for further details), and each $Z_{i,k}$ represents whether example i belongs to the k^{th} cluster (a Bernoulli variable).

Given an appropriate initial value $\boldsymbol{\theta}_0 = (\upsilon_0, \psi_0)$, the algorithm alternates between two steps : an expectation step (E-step), where we compute $t_{i,k}(\boldsymbol{\theta}_0) = \mathbb{E}[Z_{i,k}|X_1, \ldots, X_n; \boldsymbol{\theta}_0]$,

and a maximization step (M-step), where we find a vector $\boldsymbol{\theta}_1$ improving the expectation $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = \mathbb{E}[\mathcal{L}_n^c(\boldsymbol{\theta})|X_1, \ldots, X_n; \boldsymbol{\theta}_0]$ in the following sense $Q(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0) \leq Q(\boldsymbol{\theta}_0|\boldsymbol{\theta}_0)$. In the t^{th} iteration, given the last update $\boldsymbol{\theta}_{t-1} = (\upsilon_{t-1}, \psi_{t-1})$, with $\boldsymbol{\beta}_{t-1}$ the coefficients of ψ_{t-1} in the basis, the algorithm proceeds as follows:

E-step We compute the conditional expectation of the missing data $Z_{i,k}$:

$$t_{i,k}(\boldsymbol{\theta}_{t-1}) = \frac{\frac{\pi_{t-1,k}}{\sigma_{t-1,k}}\psi_{t-1}\left(\frac{X_i-\mu_{t-1,k}}{\sigma_{t-1,k}}\right)}{\sum_{l=1}^{K}\frac{\pi_{t-1,l}}{\sigma_{t-1,l}}\psi_{t-1}\left(\frac{X_i-\mu_{t-1,l}}{\sigma_{t-1,l}}\right)}.$$

M-step We compute $\boldsymbol{\theta}_t = (\upsilon_t, \psi_t)$ as follows:

$$\begin{cases} \pi_{t,k} & = \frac{1}{n}\sum_{i=1}^{n} t_{i,k}(\boldsymbol{\theta}_{t-1}), \\ (\boldsymbol{\mu}_t^T, \boldsymbol{\sigma}_t^T, \boldsymbol{\beta}_t^T)^T & = (\boldsymbol{\mu}_{t-1}^T, \boldsymbol{\sigma}_{t-1}^T, \boldsymbol{\beta}_{t-1}^T)^T - \boldsymbol{H}_{t-1}^{-1} G_{t-1}, \end{cases}$$

where $\boldsymbol{\beta}_t$ groups the coefficients of ψ_t in the truncated parameter space, G_{t-1} and $\boldsymbol{H}_{t-1}$ are, respectively, the gradient vector and the Hessian matrix of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1})$ evaluated at $\boldsymbol{\theta} = (\boldsymbol{\pi}_t^T, \boldsymbol{\mu}_{t-1}^T, \boldsymbol{\sigma}_{t-1}^T, \psi_{t-1})^T$, performing a Newton-Raphson step.

4 Numerical experiments

We illustrate the previous theoretical findings. Our simulations do not yet cover arbitrary scaling parameters. So, we perform the estimation in the case where $\boldsymbol{\sigma}^\star$ is fixed.

We consider two examples of a 2-component mixture with $\pi_1^\star = 0.4$, $\pi_2^\star = 0.6$, $\mu_1^\star = -3$ and $\mu_2^\star = 3$. In the first one, $\psi^\star$ is a Student probability density with 3 degrees of freedom. In the second one, it is a standard Laplace probability density.

The results of the simulation are in Table 1. The table shows the evolution of $\|\widehat{\psi} - \psi^\star\|_2$, $\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^\star\|_2$ and $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\star\|_2$ with respect to the sample size n. For each sample size, we run $m = 100$ experiment with different samples and aggregate the L_2 errors for the various parameters. We notice the decrease of the medians of these error distributions as the sample size increases for all the parameters.

	Student mixture			Laplace mixture		
n	$\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^\star\|_2$	$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\star\|_2$	$\|\widehat{\psi} - \psi^\star\|_2$	$\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^\star\|_2$	$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\star\|_2$	$\|\widehat{\psi} - \psi^\star\|_2$
100	$0.058_{(0.031-0.104)}$	$0.353_{(0.191-0.524)}$	$0.093_{(0.065-0.124)}$	$0.047_{(0.023-0.089)}$	$0.221_{(0.161-0.306)}$	$0.085_{(0.060-0.110)}$
200	$0.037_{(0.021-0.064)}$	$0.245_{(0.167-0.356)}$	$0.066_{(0.050-0.090)}$	$0.030_{(0.016-0.058)}$	$0.157_{(0.119-0.249)}$	$0.060_{(0.052-0.072)}$
400	$0.026_{(0.010-0.048)}$	$0.146_{(0.090-0.229)}$	$0.043_{(0.034-0.056)}$	$0.021_{(0.011-0.040)}$	$0.102_{(0.066-0.135)}$	$0.050_{(0.045-0.058)}$
800	$0.018_{(0.009-0.030)}$	$0.125_{(0.079-0.183)}$	$0.035_{(0.028-0.043)}$	$0.014_{(0.006-0.027)}$	$0.080_{(0.050-0.122)}$	$0.039_{(0.037-0.045)}$
1600	$0.014_{(0.007-0.022)}$	$0.114_{(0.080-0.172)}$	$0.026_{(0.022-0.032)}$	$0.012_{(0.004-0.019)}$	$0.056_{(0.039-0.076)}$	$0.034_{(0.030-0.038)}$

Table 1: Results of the simulation for different values of the sample size n. The Student mixture example is on the left, whereas the Laplace mixture example is on the right. Each value in the table represents the median over $m = 100$ different samples, with the interquartile range in the parentheses.

5 Conclusion and future directions

Our study addresses the semi-parametric estimation of location-scale mixture densities, extending existing approaches, which focus on the specific cases of $K = 2$ or 3 components without scale parameters. By truncating a basis in $L^2(\mathbb{R})$, we transform the problem into a finite-dimensional setting. This allows to employ M-estimation and derive consistent estimators for the finite-dimensional parameters and the coefficients of the basis representation.

Currently, we are examining the rate of convergence of the derived estimator. Using Taylor approximation around the true parameter in the truncated space and a Berry-Essen bound that depends on the dimension B (Bentkus et al. (2003)), we will be able to control the variance of the estimator $\widehat{\boldsymbol{\theta}}_{K,B,n}$, that is the quantity $\|\widehat{\boldsymbol{\theta}}_{K,B,n} - \boldsymbol{\theta}^\star_{K,B}\|_2$. Our early computations show that it converges to zero if $B = o(n^{-\frac{1}{6}})$. The rate of convergence of the bias, represented by the quantity $\|\boldsymbol{\theta}^\star_{K,B} - \boldsymbol{\theta}^\star_K\|_2$, depends on approximation theory results on the Hermite basis, where the projection errors converge to zero as B increases with known rates.

In summary, this work contributes to the theoretical understanding of semi-parametric estimation in mixture densities, offering a new framework for the estimation of mixtures in a general location-scale family.

Bibliography

Hunter, David R, Wang, Shaoli and Hettmansperger, Thomas P (2007), Inference for mixtures of symmetric distributions, *The Annals of Statistics*, , pp. 224–251.

Butucea, Cristina and Vandekerkhove, Pierre (2014), Semiparametric mixtures of symmetric distributions, *Scandinavian Journal of Statistics*, 41(1), pp. 227–239. Publisher: Wiley Online Library.

Bentkus, Vidmantas (2003), On the dependence of the Berry–Esseen bound on dimension, *Journal of Statistical Planning and Inference*, 113(2), pp. 385–402. Publisher: Elsevier.

Belomestny, Denis, Comte, Fabienne and Genon-Catalot, Valentine (2019), Sobolev-Hermite versus Sobolev nonparametric density estimation on $\mathbb{R}$, *Annals of the Institute of Statistical Mathematics*, 71(1), pp. 29–62. Publisher: Springer.

Dempster, Arthur P, Laird, Nan M and Rubin, Donald B (1977), Maximum likelihood from incomplete data via the EM algorithm, *Journal of the royal statistical society: series B (methodological)*, 39(1), pp. 1–22.

A Novel Approach to PolSAR Image Classification Using the Complex Riesz Model

Rayhan KAMMOUN [1] & Sameh KESSENTINI [2] & Raoudha ZINE [2]

[1] *Department of Mathematics, Faculty of Sciences, University of Sfax, Tunisia.*
rayhankammoun@gmail.com
[2] *Department of Mathematics, Faculty of Sciences, University of Sfax, Tunisia.*
samehkessentini@gmail.com
[3] *Department of Mathematics, Faculty of Sciences, University of Sfax, Tunisia.*
raoudha.zine@fss.usf.tn

Résumé. Ce travail propose l'utilisation de la distribution de Riesz complexe pour la classification d'images PolSAR. Les modèles traditionnels, tels que la distribution de Wishart complexe, peinent à représenter de manière adéquate la nature des données PolSAR. Nous introduisons la distribution de Riesz complexe comme une alternative plus flexible, en particulier pour les données traitées en multi-captures. L'algorithme d'Expectation-Maximization (EM) est appliqué pour estimer les paramètres de la distribution de Riesz complexe, qui est ensuite utilisée pour classer différentes régions dans les images PolSAR.

Les résultats expérimentaux montrent que la distribution de Riesz complexe surpasse le modèle de Wishart complexe en termes de précision de classification, en particulier pour les régions hétérogènes et fortement hétérogènes. La méthode a été validée sur des données réelles d'AIRSAR, réussissant à distinguer des régions de rugosité variable. Les résultats montrent que la distribution de Riesz complexe est un outil prometteur pour améliorer la classification des images PolSAR. **Mots-clés.** Distribution de Riesz complexe, classification d'images PolSAR, modèle de mélange, Expectation-Maximization, données AIRSAR.

Abstract. This work proposes the use of the Complex Riesz distribution for PolSAR image classification. Traditional models, such as the Complex Wishart distribution, struggle to represent the heterogeneous nature of PolSAR data effectively. We introduce the Complex Riesz distribution as a more flexible alternative, especially for multi-look processed data. The Expectation-Maximization (EM) algorithm is applied to estimate the parameters of the Complex Riesz distribution, which is then used to classify different regions in PolSAR images.

Experimental results demonstrate that the Complex Riesz distribution outperforms the Complex Wishart model in terms of classification accuracy, particularly for heterogeneous and highly heterogeneous regions. The method was validated on real-world AIRSAR data, successfully distinguishing regions with varying roughness. The results show that the Complex Riesz distribution is a promising tool for enhancing PolSAR image classification.

Keywords. Complex Riesz distribution, PolSAR image classification, Model mixture, Expectation-Maximization, AIRSAR data.

1 Intorduction

Polarimetric Synthetic Aperture Radar (PolSAR) imaging offers detailed structural and compositional information useful in remote sensing applications. It is widely used to monitor and analyze various environments, including forests, agricultural areas, urban landscapes, and marine surfaces. The PolSAR system measures the scattering properties of targets using different polarizations, thus providing rich information on their geometry, moisture, roughness, and composition. Applications include forest biomass estimation, crop classification, urban infrastructure monitoring, detection of natural disasters like floods and landslides, as well as studying snow and ice surfaces. However, traditional classification methods, such as the Complex Wishart distribution, often fail to capture the variability of heterogeneous and textured regions. In this context, we propose the use of the Complex Riesz distribution, which offers greater flexibility and better models the data variability in complex and textured environments.

2 PolSAR image representation

PolSAR captures the full scattering matrix at each pixel, which can be written as:

$$\mathbf{S} = \begin{pmatrix} S_{hh} & S_{hv} \\ S_{hv} & S_{vv} \end{pmatrix}.$$

By vectorizing $\mathbf{S}$ and under the assumption of scattering reciprocity ($S_{hv} = S_{vh}$) the resulting vector is

$$\mathbf{k} = \begin{pmatrix} S_{hh} & \sqrt{2}S_{hv} & S_{vv} \end{pmatrix},$$

leading to a Hermitian covariance matrix

$$\mathbf{M} = \mathbf{k}\mathbf{k}^H.$$

Multi-look processing averages multiple independent samples, forming a covariance matrix

$$\mathbf{Z} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{M}_i.$$

Since the matrix $\mathbf{Z}$ is Hermitian and positive definite, it is eligible to follow a Complex Riesz distribution on $H_r^+(\mathbb{C})$.

3 Complex Riesz distribution

3.1 Definition

The Complex Riesz distribution on the cone of positive definite Hermitian matrices $H_r^+(\mathbb{C})$ is characterized for $\Sigma \in H_r^+(\mathbb{C})$ and $\mathbf{s} = (s_1, \ldots, s_r) \in \mathbb{R}^r$, under the condition that $s_j > j-1$

for all $1 \leq j \leq r$, by the following equation:

$$\mathbb{C}\mathcal{R}_r(\mathbf{s}, \mathbf{\Sigma})\,(d\mathbf{X}) = \frac{2^{-\sum_{j=1}^{r} s_j}}{\Gamma_r\,(\mathbf{s})\,\Delta_{\mathbf{s}}(\mathbf{\Sigma})} e^{-\frac{1}{2}\langle \mathbf{\Sigma}^{-1}, \mathbf{X}\rangle} \Delta_{\mathbf{s}-r}(\mathbf{X}) \mathbf{1}_{H_r^+(\mathbb{C})}(\mathbf{X}) d\mathbf{X},$$

where $\Gamma_r(.)$ is the Complex multivariate gamma function given by

$$\Gamma_r(\mathbf{s}) = \pi^{\frac{r(r-1)}{2}} \prod_{j=1}^{r} \Gamma\left(s_j - (j-1)\right)$$

and $\Delta_{\mathbf{s}}$ the generalized power function represented by

$$\Delta_{\mathbf{s}}(\mathbf{X}) = \Delta_1(\mathbf{X})^{s_1 - s_2} \Delta_2(\mathbf{X})^{s_2 - s_3} \cdots \Delta_r(\mathbf{X})^{s_r}.$$

If $\mathbf{s} = (p, \cdots, p)$, then $\Delta_{\mathbf{s}}(\mathbf{X}) = (\det \mathbf{X})^p$.

This model generalizes the Complex Wishart distribution, better fitting PolSAR data, it is also a Riesz case when $s_1 = \ldots = s_r = p > r - 1$.

3.2 Expectation-Maximization Algorithm for mixtures of Complex Riesz distributions

To estimate the parameters of this distribution, we develop an Expectation-Maximization (EM) algorithm. The algorithm iteratively updates parameter estimates by alternating between the expectation (E) step and the maximization (M) step. We contemplate a mixture of Complex Riesz probability distributions and derive the expectation-maximization algorithm to estimate the parameters of the mixture.
Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ be N matrices drawn from one out of K Complex Riesz probability distributions with probability density function f_k with parameters $\theta_k = (\mathbf{s}_k, \mathbf{\Sigma}_k)$, $\forall 1 \leq k \leq K$. Let $(\pi_1, \pi_2, \ldots, \pi_K)$ be the vector of mixing proportions. They should belong to the interval $[0, 1]$ and sum up to one. Let $\Phi = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$, denotes the vector of unknown parameters to be estimated by the EM algorithm.

3.2.1 E-step

The expectation step is equivalent to estimate at iteration $(m+1)$ the conditional expectation $\mathbf{Q}\left(\Phi \mid \Phi^{(m)}\right)$ of the complete-data log likelihood given the previous value of parameters $\Phi^{(m)}$

$$\mathbf{Q}\left(\Phi \mid \Phi^{(m)}\right) = \sum_{i=1}^{N} \sum_{k=1}^{K} \tilde{p}_{ik} \ln\left(\pi_k^{(m)} f_k\left(\mathbf{X}_i \mid \theta_k^{(m)}\right)\right),$$

3.2.2 M-step

The maximization step of the algorithm consists of estimating, at the $(m+1)^{th}$ iteration, the new $\widehat{\Phi} = \Phi^{(m+1)}$ that maximizes the expectation subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$.

Theorem 1 *The new estimate of the unknown parameters at the maximization step of the EM algorithm is given by for* $1 \leq k \leq K$

(i)

$$\widehat{\pi}_k = \frac{1}{N} \sum_{i=1}^{N} \tilde{p}_{ik}. \tag{1}$$

(ii)

$$\widehat{s}_j^k = \begin{cases} \psi^{-1}\left(-\ln(2\Delta_1(\mathbf{\Sigma}_k)) + \dfrac{\sum_{i=1}^{N} \tilde{p}_{ik} \ln\left(\Delta_1(\mathbf{X}_i)\right)}{\sum_{i=1}^{N} \tilde{p}_{ik}} \right) & ; \quad j = 1 \\ (j-1) + \psi^{-1}\left(\ln\left(\dfrac{\Delta_{j-1}(\mathbf{\Sigma}_k)}{2\Delta_j(\mathbf{\Sigma}_k)}\right) + \dfrac{\sum_{i=1}^{N} \tilde{p}_{ik} \ln\left(\dfrac{\Delta_j(\mathbf{X}_i)}{\Delta_{j-1}(\mathbf{X}_i)}\right)}{\sum_{i=1}^{N} \tilde{p}_{ik}} \right) & ; \quad 2 \leq j \leq r \end{cases}, \tag{2}$$

where ψ *is the digamma function.*

(iii)

$$\widehat{\mathbf{\Sigma}}_k = \frac{1}{2}\left(\frac{1}{\widehat{s}_1^k} \overline{\mathbf{X}}_k + \sum_{j=1}^{r-1} \left(\frac{1}{\widehat{s}_{r-j+1}^k} - \frac{1}{\widehat{s}_{r-j}^k} \right) \left(P_j^* \left(\overline{\mathbf{X}}_k^{-1} \right) \right)^{-1} \right), \tag{3}$$

where $\overline{\mathbf{X}}_k = \dfrac{1}{\sum_{i=1}^{N} \tilde{p}_{ik}} \sum_{i=1}^{N} \tilde{p}_{ik} \mathbf{X}_i$ *for* $1 \leq k \leq K$.

These iterations continue until convergence, ensuring robust parameter estimation.

4 Application to Polarimetric SAR data

We apply the developed theoretical results to real PolSAR data in order to assess the performance of the complex Riesz model in a practical setting.

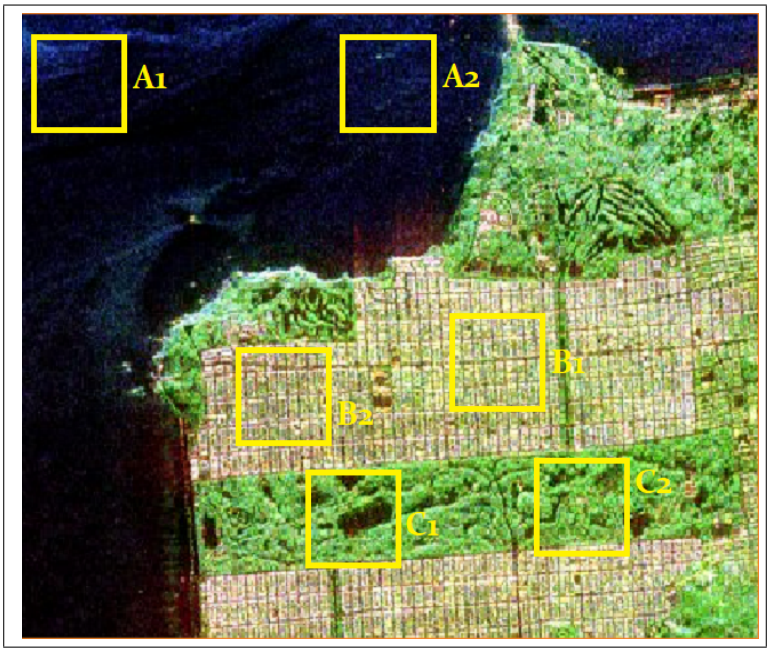


Figure 1: AIRSAR image (HH data)

Consider six areas in Figure 1, each of size 80×80 pixels, with three degrees of roughness:

- Areas A_1 and A_2: homogeneous
- Areas B_1 and B_2: heterogeneous
- Areas C_1 and C_2: extremely heterogeneous

The classification of the combined regions A_1, A_2, B_1, B_2, C_1, and C_2, covering a total of 38,400 pixels, is performed using the Expectation-Maximization algorithm under the Complex Riesz (EMCR) model.

Table 1: Parameters estimated by Expectation-Maximization algorithm.

	Class 1
$\widehat{\pi}_1$	0.33319
$\widehat{\mathbf{s}}_1$	(2.2088, 4.3756, 4.1408)
$\widehat{\mathbf{\Sigma}}_1$	$\begin{pmatrix} 8.2114 \times 10^{-4} & & \\ 3.1653 \times 10^{-5} + 1.3052i \times 10^{-4} & 7.2345 \times 10^{-5} & \\ 1.6371 \times 10^{-3} - 1.9199i \times 10^{-4} & 2.2505 \times 10^{-5} - 2.9471i \times 10^{-4} & 3.6376 \times 10^{-3} \end{pmatrix}$
$\det\left(\widehat{\mathbf{\Sigma}}_1\right)$	1.2910×10^{-11}
	Class 2
$\widehat{\pi}_2$	0.48874
$\widehat{\mathbf{s}}_2$	(1.6359, 2.6949, 3.3779)
$\widehat{\mathbf{\Sigma}}_2$	$\begin{pmatrix} 1.8737 \times 10^{-2} & & \\ 2.6503 \times 10^{-3} + 6.6586i \times 10^{-4} & 5.0046 \times 10^{-3} & \\ 6.0325 \times 10^{-4} - 1.0855i \times 10^{-3} & -1.9340 \times 10^{-4} - 7.1490i \times 10^{-4} & 8.7486 \times 10^{-3} \end{pmatrix}$
$\det\left(\widehat{\mathbf{\Sigma}}_2\right)$	7.4141×10^{-7}
	Class 3
$\widehat{\pi}_3$	0.17807
$\widehat{\mathbf{s}}_3$	(0.9113, 2.6562, 3.1089)
$\widehat{\mathbf{\Sigma}}_3$	$\begin{pmatrix} 2.2530 \times 10^{-1} & & \\ 5.7442 \times 10^{-2} - 1.0045i \times 10^{-3} & 2.4833 \times 10^{-2} & \\ -6.0101 \times 10^{-2} + 2.5773i \times 10^{-3} & -1.8590 \times 10^{-2} - 5.3166i \times 10^{-3} & 6.9215 \times 10^{-2} \end{pmatrix}$
$\det\left(\widehat{\mathbf{\Sigma}}_3\right)$	1.1223×10^{-4}

Table 1 presents the scale and shape parameters of the Complex Riesz distributions for three classes. The first class, corresponding to homogeneous pixels with the smallest covariance determinant, describes the A_1 and A_2 areas with proportion $\widehat{\pi}_1 = 0.333$. The second class, covering slightly heterogeneous pixels ($\det(\widehat{\mathbf{\Sigma}}2) = 7.4 \times 10^{-7}$), constitutes nearly half of the mixture. The third class, representing highly heterogeneous pixels, accounts for $\widehat{\pi}_3 = 0.178$ of the sample.

5 Conclusion

In conclusion, this study establishes the Complex Riesz distribution as a powerful tool for PolSAR image classification. By extending the EM algorithm we provide a robust framework for analyzing PolSAR data.

Bibliographie

Kammoun, R., Kessentini, S. et Zine, R. (2025), Classification of PolSAR data with the Complex Riesz distribution, *Communications in Statistics-Theory and Methods*, pp. 1-28.

Hassairi, A., Lajmi, S. et Zine, R. (2008), A characterization of the Riesz probability distribution, *Journal of Theoretical Probability*, 21, pp. 773-790.

Kessentini, S., Tounsi, M. et Zine, R. (2018), The Riesz probability distribution: generation and EM algorithm, *Communications in Statistics-Simulation and Computation*, 49(8), pp. 2114-2133.

Carvalho, N. C. R. L., Sant'Anna Bins, L. et Siqueira Sant'Anna, S. J. (2019), Analysis of stochastic distances and Wishart mixture models applied on PolSAR images, *Remote Sensing*, 11(24), pp. 2994.

Muirhead, R. J. (2009), *Aspects of multivariate statistical theory*, John Wiley & Sons.

Pottier, E. (2010), Recent advances in the development of the open source toolbox for polarimetric and interferometric polarimetric SAR data processing: the PolSARpro v4.1.5 software, *2010 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2527-2530.

Carvalho, N. C. R. L., Sant'Anna Bins, L. et Siqueira Sant'Anna, S. J. (2019), Analysis of stochastic distances and Wishart mixture models applied on PolSAR images, *Remote Sensing*, 11(24), pp. 2994.

Classification des données d'observance d'un traitement de l'apnée du sommeil à l'aide d'un modèle de mélange de régression basé sur une chaîne de Markov

Kajal Eybpoosh [1] & Elodie Vernet [1] & Sebastien Bailly [2] & Jean-Louis Pépin [2] & Adeline Samson [1] & Vincent Brault [1]

[1] *Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France ;*
[2] *Univ. Grenoble Alpes, INSERM, HP2, 38000 Grenoble, France ;*
elodie-edith.vernet@univ-grenoble-alpes.fr ;
vincent.brault@univ-grenoble-alpes.fr

Résumé. Le but est d'identifier différents types de profils d'utilisateurs de la ventilation à pression positive continue (PPC) qui est utilisé par les patients souffrant d'apnée obstructive du sommeil (AOS). Deux cent vingt-deux individus ont été équipés d'une PPC prescrite au CHU de Grenoble entre 2016 et 2018. Les profils d'adhésion au traitement sont définis à partir des données de télémédecine, à savoir l'utilisation nocturne de la PPC. Ces séries temporelles contiennent un excès de zéros pour la non-utilisation du dispositif ou le dysfonctionnement de la transmission des données. Nous proposons donc une méthode pour regrouper les profils d'adhésion en fonction de leur comportement temporel et en tenant compte de l'excès de zéros. La méthode associe un modèle de mélange et des chaînes de Markov dont la loi conditionnelle est soit une loi gaussienne soit un Dirac en zéro. Nous développons la méthode d'estimation avec un algorithme EM et discutons de la sélection du nombre de groupes par BIC qui a une forme particulière. Les propriétés sont illustrées sur données simulées.

Mots-clés. Classification non supervisée, Modèle de Markov, Excès de zeros, Mélange de régression, Apnée du sommeil, données de la thérapie PPC

Abstract. The objective is to identify several types of user profiles of continuous positive airway pressure (CPAP) used by patients suffering Obstructive Sleep Apnea (OSA). Two hundred and twenty two subjects were equipped with CPAP prescribed at the Grenoble University Hospital between 2016 and 2018. Patterns of adherence are defined from telemedicine data, namely their nightly CPAP use. These adherence time series contain an excess of zeros for non-use of the device or data transmission malfunction. We thus propose a method for clustering the adherence profiles according to their temporal behavior along day taking into account the excess of zeros. The method combines a mixture model and Markov chains where the conditional distribution is either Gaussian or a Dirac delta distribution in zero. We develop the estimation method with an EM algorithm and discuss the selection of the number of clusters by BIC which is specific for this model. The properties of the estimators are illustrated on simulated data.

Keywords. Clustering, Markov Model, Excess of zeros, Mixture of regression, Sleep apnea, CPAP device data

1 Introduction

Un des traitements des patients souffrant d'apnée obstructive du sommeil (AOS) est l'utilisation de la ventilation à pression positive continue (PPC). Ce traitement permet d'éviter nombres d'effets secondaires (hypertension artérielle, insuffisance respiratoire, etc). Cependant, ce traitement est très lourd et beaucoup de patients l'interrompent. Les médecins s'intéressent à identifier différents profils d'adhérence au traitement parmi la population de patients, à partir de mesures quotidiennes d'utilisation de l'appareil PPC.

L'enjeu statistique est de proposer une classification de séries temporelles comportant un nombre important de valeurs nulles (correspondant aux absences d'utilisation) et de valeurs positives. Nous introduisons un modèle statistique de classification sous la forme d'un modèle de mélange de régression. Chaque modèle de régression est défini par une chaine de Markov latente décrivant la présence de zeros et un modèle de régression linéaire lorsque le patient utilise l'appareil. Ce nouveau modèle pose un certain nombre de questions statistiques, auxquelles nous répondons d'un point de vue théorique : identifiabilité, estimation, consistance, sélection du nombre de composantes dans le modèle de mélange. Enfin, nous illustrerons ces propriétés sur des données simulées.

2 Les données d'observance

Le jeu de données utilisé dans ce papier a été étudié par Bottaz-Bosson et al (2021). Il inclut $n = 222$ patients qui ont commencé un traitement PPC. La variable d'intérêt est le nombre d'heures d'utilisation de l'appareil PPC par nuit durant $N = 91$ nuits. La figure 1 montre les séries temporelles de l'observance de la PPC pour 6 individus.

3 Le modèle statistique de mélange

On note $Y_{i,j}$ le nombre d'heures d'utilisation de l'appareil pendant la nuit j par l'individu i, pour $j = 1, \ldots, N$ et $i = 1, \ldots, n$.

On suppose que les individus sont regroupés dans K clusters décrivant les différents profils d'observance. On note $Z_{i,k}$, pour $k = 1, \ldots, K$, la variable binaire indiquant si l'individu i appartient au cluster k et $P_k = P(Z_{i,k} = 1)$ la probabilité d'appartenir au cluster k.

Considérons un individu i appartenant au cluster k. Son nombre d'heures $Y_{i,j}$ d'utilisation est strictement positif si l'appareil est utilisé et nul sinon. Nous introduisons un vecteur binaire latent S_i, de taille N, décrivant si l'appareil a été utilisé ou non. On suppose que le processus $S_i = (S_{i,j}, j = 1, \ldots, N)$ est une chaîne de Markov décrite par sa loi initiale :

$$P(S_{i,j} = 1 | Z_{i,k} = 1) = \pi_k$$

et sa matrice de transition, pour $l, l' = 0, 1$,

$$P(S_{i,j} = l' | S_{i,j-1} = l; Z_{i,k} = 1) = A_{k;l,l'}.$$

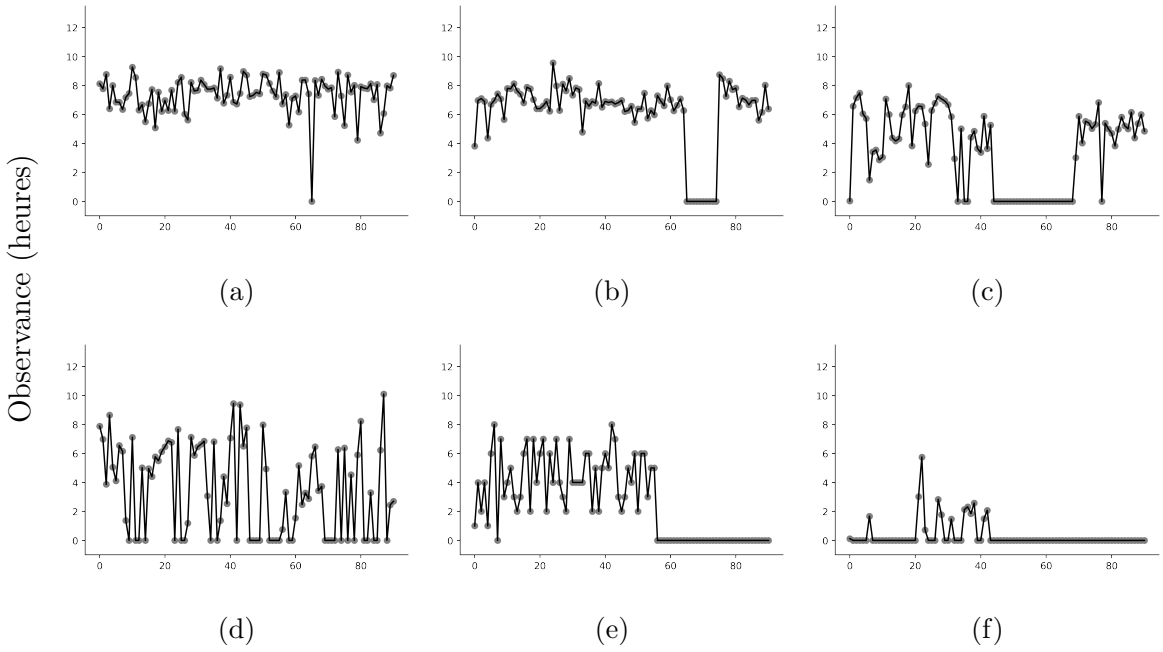


Figure 1: Séries temporelles de l'observance de la PPC durant les 91 premières nuits de 6 patients.

On introduit maintenant le modèle décrivant la dynamique du nombre d'heures $Y_{i,j}$. Lorsque l'appareil n'est pas utilisé ($S_{i,j} = 0$), $Y_{i,j} = 0$. Sinon, c'est-à-dire si $S_{i,j} = 1$, et si l'individu est dans le cluster k, i.e. $Z_{i,k} = 1$; alors $Y_{i,j}$ suit un modèle linéaire (en temps) d'ordonnée à l'origine a_k et de pente b_k :

$$Y_{i,j}|S_{i,j} = 1 \text{ et } Z_{i,k} = 1; = a_k + jb_k + \epsilon_{i,j,k} \quad \text{avec } \epsilon_{i,j,k} \sim \mathcal{N}(0, \sigma_k^2).$$

Ainsi les paramètres inconnus du modèle sont $\mathbf{P} = (P_1, \ldots, P_K)$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\mathbf{A} = (A_1, \ldots, A_K)$ et $\boldsymbol{\theta}_k = (a_k, b_k, \sigma_k)$ et on note $\boldsymbol{\theta} = (\mathbf{P}, \boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$.

On peut remarquer que les chaînes de Markov S_i non observés peuvent être facilement retrouvées comme suit :

$$S_{i,j} = 1_{Y_{ij} \neq 0}.$$

Le premier résultat que nous obtenons est l'identifiabilité du modèle, qui n'est pas automatique en raison du modèle de mélange.

Théorème 3.1 (Identifiabilité) *On suppose que*

(C1) pour tout $1 \leq k \leq K$,
$P_k > 0$, $0 < \pi_k < 1$, $A_{k;1,1} > 0$, $A_{k;1,0} > 0$, $A_{k;0,0} > 0$, $A_{k;0,1} > 0$,

(C2) $N \geq 3$ *et les* σ_k^2 *sont distincts,*

alors le modèle est identifiable à la re-labellisation des clusters près.

4 Estimation des paramètres et sélection du nombre de clusters

On considère tout d'abord un nombre de clusters K fixé. L'estimateur du maximum de vraisemblance n'est pas explicite et on utilise l'algorithme espérance-maximisation (EM) définit comme suit pour l'approcher.

- **Initialisation** : Choisir une partition $z^{(0)}$ et faire l'étape M pour obtenir $\boldsymbol{\theta}^{(0)}$.
- **Itération c, étant donnée la valeur actuelle $\boldsymbol{\theta}^{(c)}$:**
  - **Étape E :** Calculer les probabilités $r_{i,k}^{(c)}$ avec

$$r_{i,k}^{(c)} = \frac{P_k^{(c)}\left(\pi_k^{(c)}\right)^{S_{i,1}}\left(1-\pi_k^{(c)}\right)^{1-S_{i,1}}\left(\prod_{j=2}^{N} A_{k;S_{i,j-1},S_{i,j}}^{(c)}\right)\left[\prod_{j=1}^{N}\varphi\left(Y_{i,j};a_k^{(c)}+jb_k^{(c)},\sigma_k^{(c)}\right)^{S_{i,j}}\right]}{\sum_{k'=1}^{K} P_{k'}^{(c)}\left(\pi_{k'}^{(c)}\right)^{S_{i,1}}\left(1-\pi_{k'}^{(c)}\right)^{1-S_{i,1}}\left(\prod_{j=2}^{N} A_{k';S_{i,j-1},S_{i,j}}^{(c)}\right)\left[\prod_{j=1}^{N}\varphi\left(Y_{i,j};a_{k'}^{(c)}+jb_{k'}^{(c)},\sigma_{k'}^{(c)}\right)^{S_{i,j}}\right]},$$

    où $\varphi\left(Y_{i,j};a_k^{(c)}+jb_k^{(c)},\sigma_k^{(c)}\right)$ est la densité gaussienne en $Y_{i,j}$ de moyenne $a_k^{(c)}+jb_k^{(c)}$ et d'écart-type $\sigma_k^{(c)}$.
  - **Étape M :**

$$P_k^{(c+1)} = \frac{\sum_{i=1}^{n} r_{i,k}^{(c)}}{n}, \quad \pi_k^{(c+1)} = \frac{\sum_{i=1}^{n} r_{i,k}^{(c)} S_{i,1}}{\sum_{i=1}^{n} r_{i,k}^{(c)}},$$

$$A_{k;0,1}^{(c+1)} = \frac{\sum_{i=1}^{n}\sum_{j=2}^{N} r_{i,k}^{(c)}(1-S_{i,j-1})S_{i,j}}{\sum_{i=1}^{n} r_{i,k}^{(c)}\left(N-1-\sum_{j=1}^{N-1} S_{i,j}\right)}, \quad A_{k;1,1}^{(c+1)} = \frac{\sum_{i=1}^{n}\sum_{j=2}^{N} r_{i,k}^{(c)} S_{i,j-1}S_{i,j}}{\sum_{i=1}^{n} r_{i,k}^{(c)}\sum_{j=1}^{N-1} S_{i,j}},$$

$$a_k^{(c+1)} = \frac{\left(\sum_{i=1}^{n} r_{i,k}^{(c)}\sum_{j=1}^{N} j^2 S_{i,j}\right)\times\left(\sum_{i'=1}^{n} r_{i',k}^{(c)}\sum_{j'=1}^{N} S_{i',j'}Y_{i',j'}\right)-\left(\sum_{i=1}^{n} r_{i,k}^{(c)}\sum_{j=1}^{N} jS_{i,j}\right)\times\left(\sum_{i'=1}^{n} r_{i',k}^{(c)}\sum_{j'=1}^{N} S_{i',j'}j'Y_{i',j'}\right)}{\left(\sum_{i=1}^{n} r_{i,k}^{(c)}\sum_{j=1}^{N} S_{i,j}\right)\times\left(\sum_{i'=1}^{n} r_{i',k}^{(c)}\sum_{j'=1}^{N} j'^2 S_{i',j'}\right)-\left(\sum_{i=1}^{n} r_{i,k}^{(c)}\sum_{j=1}^{N} jS_{i,j}\right)^2},$$

$$b_k^{(c+1)} = \frac{\left(\sum_{i=1}^{n} r_{i,k}^{(c)}\sum_{j=1}^{N} S_{i,j}\right)\times\left(\sum_{i'=1}^{n} r_{i',k}^{(c)}\sum_{j'=1}^{N} j' S_{i',j'}Y_{i',j'}\right)-\left(\sum_{i=1}^{n} r_{i,k}^{(c)}\sum_{j=1}^{N} jS_{i,j}\right)\times\left(\sum_{i'=1}^{n} r_{i',k}^{(c)}\sum_{j'=1}^{N} S_{i',j'}Y_{i',j'}\right)}{\left(\sum_{i=1}^{n} r_{i,k}^{(c)}\sum_{j=1}^{N} S_{i,j}\right)\times\left(\sum_{i'=1}^{n} r_{i',k}^{(c)}\sum_{j'=1}^{N} j'^2 S_{i',j'}\right)-\left(\sum_{i=1}^{n} r_{i,k}^{(c)}\sum_{j=1}^{N} jS_{i,j}\right)^2},$$

$$\text{et} \quad \sigma_k^{2(c+1)} = \frac{\sum_{i=1}^{n} r_{i,k}^{(c)}\sum_{j=1}^{N} S_{i,j}(Y_{i,j}-a_k^{(c+1)}-j.b_k^{(c+1)})^2}{\sum_{i=1}^{n} r_{i,k}^{(c)}\sum_{j=1}^{N} S_{i,j}}.$$

Lorsque $N \geq 3$ est fixé, on peut prouver la consistance de l'estimateur du maximum de vraisemblance sur un compact à la re-labellisation des clusters près.

Pour sélectionner le nombre optimal K de clusters, on utilise le critère d'information Bayésien (BIC) introduit par Schwarz (1978). La démarche de Lebarbier et Mary-Huard (2006) donne le critère suivant.

Théorème 4.1 (BIC) *Supposons qu'il existe $c > 0$ tel que pour tout $k \in \{1, \ldots, K\}$ et $\ell, \ell' \in \{0, 1\}$, nous avons $\pi_k \geq c$, $\pi_{k,\ell} \geq c$ and $A_{k,\ell\ell'} \geq c$,. Le critère BIC est le suivant :*

$$BIC(K) = -\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_K} \log p(Y, S; \boldsymbol{\theta}) + \frac{7K - 1}{2} \log(n) + 5K \log(N).$$

avec $\boldsymbol{\Theta}_K$ l'espace des paramètres lorsque le nombre de clusters est égal à K et $p(Y, S; \boldsymbol{\theta})$ la vraisemblance.

5 Simulations

On simule des données selon le modèle décrit précédemment et selon le même modèle mais avec un bruit de Laplace au lieu de Gaussien avec des paramètres quelconques (scenario 1) et avec des paramètres estimés avec les données réels lorsqu'on choisit 3 clusters (scenario 2). Le vrai nombre de clusters est de 3. Le but est de comparer deux critères de sélection du nombre de clusters. Pour le premier on minimise :

$$CRIT(K) = -\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_K} \log p(Y, S; \boldsymbol{\theta}) + \frac{7K - 1}{2} \log(n)$$

Le second critère minimise le critère BIC du Théorème 4.1. On maximise la vraisemblance avec l'algorithme EM décrit dans la partie 4. On initialise celui-ci une fois, dix fois ou vingt-cinq fois.

Les résultats de la sélection de modèle sont visibles dans la Figure 2. Dans le cas Gaussien, les deux critères sélectionnent le bon nombre de clusters (3) dans les différents scenari et pour tous les nombres d'initialisations. Dans le cas du bruit de Laplace, le critère BIC a tendance à mieux sélectionner le nombre de clusters dans tous les scenari et pour tous les nombres d'initialisations.

Bibliographie

Bottaz-Bosson G., Hamon A., Pépin J-L, Bailly S. et Samson A. (2021) Continuous positive airway pressure adherence trajectories in sleep apnea: Clustering with summed discrete Fréchet and dynamic time warping dissimilarities, *Statistics in Medicine* 40(24): 5373–5396.

Lebarbier É et Mary-Huard T. (2006) Une introduction au critère BIC: fondements théoriques et interprétation. *Journal de la Société française de statistique* 147(1): 39–57.

Schwarz G. (1978) Estimating the dimension of a model. *Annals of Statistics* 6: 461-464.

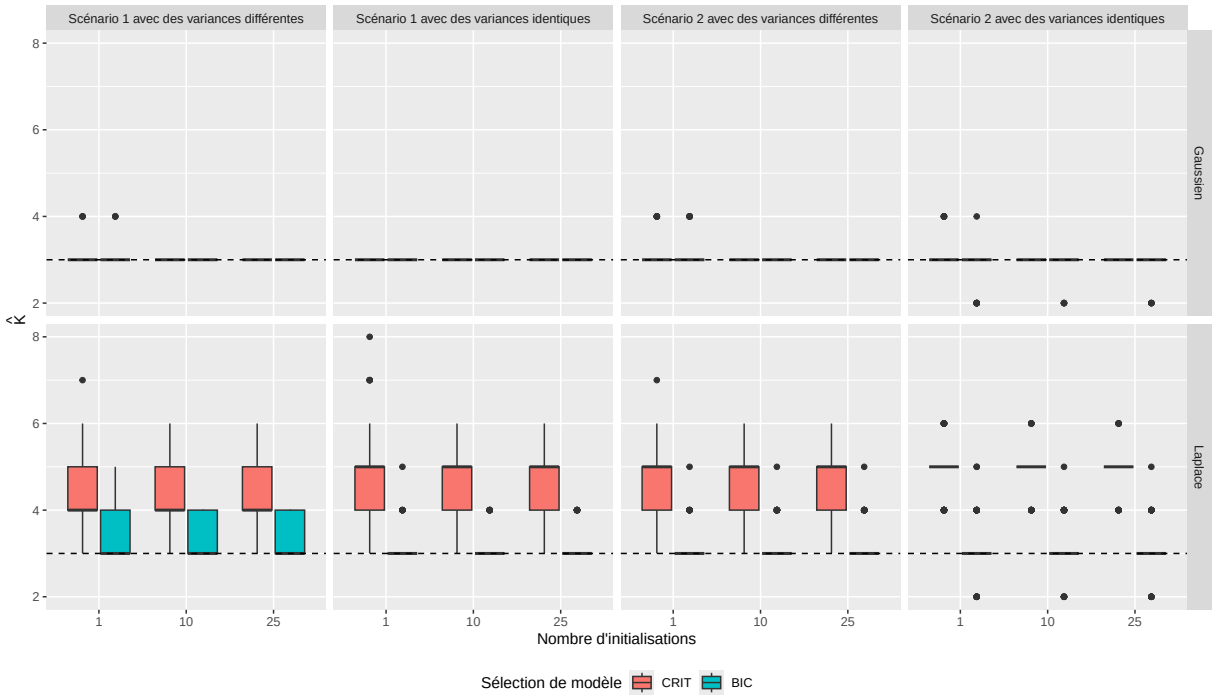


Figure 2: Estimation du nombre de clusters selon 2 scenari avec des variances identiques et différentes (colonnes) et selon deux bruits (lignes) en fonction du nombre d'initialisations (axe des abscisses) et du critère de sélection (couleurs).

Prix actuariat + Session Groupe Risques AEF

Modélisation et prédiction du Best Estimate par réseau de neurones génératifs

Kouassi Alphonse Nathan Ebrotie*[1]

[1]EURIA – Prix de Statistiques pour l'Actuariat 2025 – France

Résumé

La réglementation Solvabilité ii (Sii) impose aux assureurs un certain nombre de contraintes dont l'évaluation de leurs provisions techniques de la façon la plus juste possible, ces provisions ayant une composante dite Best Estimate (BE). Un autre enjeu de Sii, concerne l'estimation prospective de certaines métriques comme le ratio de Solvabilité dans le cadre de l'ORSA ou pour des allocations stratégiques d'actifs. Afin de respecter ces exigences du régulateur, les assureurs utilisent généralement des modèles d'Asset and Liability Management (ALM) afin d'effectuer des projections des différents flux financiers en fonction de scénarios économiques. À partir de ces projections, les différents éléments d'intérêts (BE, ratio, etc.) peuvent être obtenus en utilisant des méthodes basées sur des simulations comme l'approche classique de Simulations dans les Simulations (SdS). Toutefois, en pratique, cette méthode fondée sur le Monte-Carlo est très couteuse en temps de calcul. C'est pour cette raison que des techniques de réduction de temps de calcul ont été développées. Dans la littérature actuarielle, plusieurs mémoires se sont attardés sur la mise en place de ces métamodèles pour des fins identiques à ceux évoqués plus haut, mais adoptent des approches alternatives qui ne se généralisent pas souvent et ne permettent pas parfois de capturer efficacement les distributions temporelles des éléments d'intérêts, sachant que cette distribution peut être utile pour des problématiques d'estimations prospectives. Avec l'avènement d'outils comme ChatGPT, DALL-E, Midjourney, les modèles génératifs font parler d'eux de façon spectaculaire, et laissent entrevoir de nouvelles opportunités en termes de modélisation au vu des capacités des modèles sous-jacents. C'est dans la perspective de tirer profit des possibilités qu'offrent ces modèles génératifs dans des tâches de modélisation que ce mémoire s'inscrit. Notre étude se focalise sur l'utilisation des modèles génératifs afin d'approximer le Best Estimate. Cette estimation s'intègre également dans un cadre prospectif et peut permettre de réaliser des anticipations aussi bien sur le BE que sur l'évolution de métriques comme le SCR et le ratio de Solvabilité.

Mots-Clés: Solvabilité ii, Best Estimate, proxy, réseaux de neurones génératifs.

*Intervenant

Efficient estimation for EV regression models of tail risks

Julien Hambuckers
HEC Liège, University of Liège, Belgium
Marie Kratz
ESSEC Business School, CREAR,France
Antoine Usseglio-Carleve,
Avignon Université, LMA, France,

Extreme value regression (EVR), a method introduced by Davison and Smith (1990), is a crucial tool for analyzing the factors influencing extreme losses, for instance of financial entities. Using asymptotic theorems from Extreme Value Theory, EVR models assume that the tail distribution of a variable of interest can be approximated by a generalized Pareto distribution (GPD). However, heterogenous data are difficult to model, hence EVR models become complicated by the heterogeneity of the data, and the tail parameters (particularly the tail index) are often treated as functions of covariates that influence the occurrence of extreme events. Practically, the estimation of EVR models typically involves the peaks-over-threshold (POT) method, which begins by selecting a high threshold, discarding observations smaller than this threshold, and using the remaining data (the extreme values) to estimate the tail parameters. As usual, the selection of an appropriate threshold is a challenge, in particular for EVR models; this becomes specifically true when applied, e.g., to hedge fund returns. The POT approach often discards a large portion of the data (up to 95%), leading to a significant loss of statistical power, especially when the dataset is small. Considering the example of hedge funds, as the number of hedge funds reporting monthly returns is relatively small compared to the number of stocks, applying a two-step estimation approach can cause substantial inefficiencies. Moreover, methods that use global or regression-based thresholds, such as empirical quantiles or conditional quantiles, further exacerbate this issue by causing significant loss of efficiency in the estimation stage.

In light of these challenges, we propose a novel estimation method that improves the estimation of tail risk in EVR models by overcoming the inefficiencies of traditional threshold selection. We use an auxiliary splicing distributional regression model, based on a hybrid model suggested in the

non-regression setting by Debbabi et al. (2017) (see also Dacorogna et al., 2023), which extends the EVR model below the extreme value threshold. This hybrid model combines a distribution for the body of the data (e.g., Gaussian) and the GPD for the upper tail, with an exponential bridge between the two components. This allows for more efficient estimation without the need of selecting an arbitrary initial tail-threshold. The junction point between the exponential bridge and the tail of the distribution can be interpreted as a conditional threshold in an EVR context, which allows for a more accurate estimation of the tail regression parameters.

To further address potential misspecifications, the authors propose a robust estimation strategy based on a censored likelihood principle. This approach mitigates the bias that may arise from incorrectly specified models for the lower tail (or body) of the distribution. By using the full dataset, including data from the body of the distribution, the method increases the efficiency of the estimates for the tail parameters compared to classical POT-based methods.

The authors apply this new methodology to a dataset of 1,982 hedge funds, spanning 20 years, to study the relationship between financial conditions and hedge fund tail risk. Their findings align with the funding liquidity theory of Brunnermeier and Pedersen (2009), showing that hedge fund tail risk is significantly influenced by market factors such as financial uncertainty, stock market performance, and funding conditions. The study also demonstrates that the proposed method provides better predictive power for abnormal hedge fund returns than traditional methods.

The authors emphasize the superiority of their approach, noting that it provides a more efficient and robust method for estimating tail regression models, especially when data is limited. This is made possible by avoiding the need to discard large portions of the data. They also highlight the practical implications of their work for understanding hedge fund risk, especially in the context of the broader financial system, where extreme events have the potential to trigger significant instability. Future work will explore the out-of-sample predictive capabilities of the tail risk measure.

To conclude and summarize, in this study, a more efficient and robust framework is developed for studying tail risk, with an application to hedge funds; this can be a helpful tool for (financial) risk management, specifically when facing heterogenous data. We propose:

(i) A flexible and dynamic EVR model, which outperforms traditional POT-based methods (e.g. Chavez-Demoulin et al. (2016); Mhalla et al. (2022)) and the semi-parametric approach of Kelly and Jiang (2014) in terms of explanatory power for hedge fund abnormal returns

(ii) An improved Estimation Method, using information from the bulk of the distribution thanks to a splicing (hybrid) distributional regression model that allows for simultaneous estimation of the tail regression parameters and the threshold, addressing inefficiencies inherent in traditional POT methods

(iii) A robust Estimation Strategy, with a procedure based on censored maximum likelihood, which helps mitigate the impact of misspecifications in the lower tail of the distribution

(iv) An empirical application to Hedge Funds: Using data from 1,982 hedge funds over 20 years, we show the effectiveness of the new approach in estimating tail risk and identifying its dependence on financial conditions

Keywords
Extreme value theory; generalized Pareto regression; censored maximum likelihood; hybrid distribution

Systemic Risk Conditional On Comovement

Yanxi Hou* Haitao Huang† Liang Peng‡

February 13, 2025

Systemic risk has emerged as a critical concern for both the financial industry and insurance business since the 2008 global financial crisis, which has stimulated a plethora of research into measuring, modeling, and inferring systemic risk. A prevalent systemic risk measure is the so-called CoVaR introduced by Adrian and Brunnermeier (2016), which is defined as the conditional quantile of system loss at risk level $q \in (0, 1)$ given certain predictors and an event that an individual loss equals its Value-at-Risk (VaR) at the same risk level q. When the given predictors are indispensable, nonparametric inference for CoVaR is rather inefficient. Since quantile regression has been proven effective in modeling and inferring the conditional quantile of a loss given certain predictors, Adrian and Brunnermeier (2016) further employ two quantile regressions to model and infer CoVaR: one quantile regression for the individual loss given some macroeconomic predictors and the other quantile regression for the system loss given macroeconomic predictors and the individual loss. The theoretical aspects of the two-quantile regression model have recently been studied by Leng, He, Hou and Peng (2024). The academic literature abounds with diverse applications and interpretations of systemic risk measures. For example, one generalization of CoVaR is to modify the conditional event from loss being at its VaR to loss being above its VaR. Another extension replaces the conditional VaR of the response by the conditional Expected Shortfall (ES), culminating in the CoES systemic risk measure. In portfolio selection, systemic risk is often defined as the conditional VaR or ES of the portfolio given systemic losses, rather than individual losses at or above their VaR.

On the other hand, (extreme) comovement across financial markets, institutions, and assets has long been pivotal to risk management, predating the extensive research on systemic risk measures. Comovement is typically characterized as the conditional probability of a substantial loss in one market (or asset) given a significant loss in another market (or asset), without including additional predictors. Numerous studies have been dedicated to modeling comovement and obtaining significant empirical evidence. In the context of financial market comovement, seminal works have provided foundational insights into the dynamics of financial market correlations under stress. In examining comovement within housing markets, alleged to have exacerbated the housing crisis, results in the literature have shed light on the interconnectedness and spillover effects within the housing sector, which are crucial for understanding their broader economic implications.

Since systemic risk quantifies the impact of an individual on the system, while comovement

*School of Data Science, Fudan University, China.
†School of Digital Economics and Management, Nanjing University, China
‡Maurice R. Greenberg School of Risk Science, Georgia State University, USA.

measures the possibility of simultaneous substantial losses in two markets or assets, a practical question is how to measure the impact of the comovement between two institutions on the system, a concept we refer to as *Systemic Risk conditional on Comovement*. Recently, Bei, Lin and Zhou (2024) introduce a measure that defines this as the conditional Value-at-Risk (VaR) of the systemic loss, given that both institutions have exceeded their respective VaRs, without the use of any additional predictors. Hence, a nonparametric inference is feasible for this defined measure.

In this paper, we follow and generalize the CoVaR definition in Adrian and Brunnermeier (2016) to define a systemic risk measure conditional on comovement as the conditional VaR of the system loss given certain predictors and a comovement event with zero probability and depending on predictors. Specifically, we define the comovement event with zero probability as the event when both individual losses are above or equal to their respective conditional VaRs given the predictors, yet at least one loss is precisely at its conditional VaR given the predictors. Consequently, the comovement event depends on the predictors, too. We denote this measure as CoVaRCM, which differs from the measure in Bei, Lin and Zhou (2024) in two respects: it permits the inclusion of predictors, and the comovement event is not only of zero probability but also depends on these predictors. Hence, the modeling and inference method employed by Bei, Lin and Zhou (2024) is inapplicable to our proposed CoVaRCM.

This paper makes two primary contributions. The first contribution extends the two-quantile regression model presented in Adrian and Brunnermeier (2016) to a three-quantile regression framework for the inference of CoVaRCM. This new framework contains: the first quantile regression for the loss of the ith institution given predictors, the second quantile regression for the loss of the jth institution given predictors, and the third quantile regression for the systemic loss given both predictors and the comovement event. To infer CoVaRCM, we employ quantile regression estimation techniques. Additionally, we introduce two novel metrics along with a bootstrap inference procedure designed to assess the influence of comovement on systemic risk in comparison to CoVaR, which does not account for comovement.

The second contribution empirically shows the importance of accounting for comovement in monitoring systemic risk by revisiting the dataset analyzed by Adrian and Brunnermeier (2016) to examine the impact of comovement on systemic risk. Given that our comparative metrics are predicated on predictors, we categorize paired institutions into having negative, zero, and positive impacts based on a bootstrapped confidence interval for the comparative metrics for each observed predictor vector. Ultimately, we determine the ratios of these negative, zero, and positive comparative metrics across predictor vectors. Our analysis reveals the comovement effect on the systemic risk and the dynamics of such impact, which depend on both the predictors and the paired institutions. Consequently, by leveraging predictors, we can not only forecast our risk measure but also identify "critical" institution pairs through the proposed comparative metrics, advocating for the monitoring of these pairs above and beyond the "important" individuals typically detected by conventional systemic risk measures such as CoVaR.

Etude asymptotique des statistiques de cluster

Gloria Buriticá [1] & Olivier Wintenberger [2]

[1] *AgroParisTech - INRAE - Campus Agro Paris-Saclay, 22 place de lAgronomie, 91120 Palaiseau, gloriapatricia.buriticaborda@agroparistech.fr*
[2] *LPSM, Sorbonne Universit 4 place Jussieu 75005 Paris, FRANCE, olivier.wintenberger@sorbonne-universite.fr*

Résumé. Les extrêmes se manifestent dans les séries temporelles stationnaires à variation régulière sous forme de courtes périodes comportant plusieurs observations de grande amplitude, appelées blocs extrêmaux. Nous étudions des statistiques de cluster qui résument le comportement de fonctions agissant sur ces blocs extrêmaux. Parmi les exemples de statistiques de cluster, on trouve l'indice extrêmal et les probabilités de taille de cluster. L'objectif de notre travail est double. Premièrement, nous établissons la normalité asymptotique des estimateurs par blocs pour l'inférence sur les clusters, basés sur des observations consécutives de grandes normes ℓ^α, où α est l'indice de variation régulière. Deuxièmement, nous vérifions les conditions requises sur des modèles classiques tels que les modèles linéaires et les solutions d'équations de récurrence stochastiques. Concernant les modèles linéaires, nous prouvons que la variance asymptotique des estimateurs classiques fondés sur les indices de cluster est nulle, comme conjecturé initialement dans Hsing (1993).

Mots-clés. Théorie des valeurs extrêmes; Séries temporelles.

Abstract. Extremes occur in stationary regularly varying time series as short periods with several large observations, known as extremal blocks. We study cluster statistics summarizing the behavior of functions acting on these extremal blocks. Examples of cluster statistics are the extremal index, cluster size probabilities, and other cluster indices. The purpose of our work is twofold. First, we state the asymptotic normality of block estimators for cluster inference based on consecutive observations with large ℓ^α-norms, where $\alpha > 0$ is the index of regular variation. Second, we verify the conditions we require on classic models such as linear models and solutions of stochastic recurrence equations. Regarding linear models, we prove that the asymptotic variance of classical index cluster-based estimators is null as first conjectured in Hsing (1993).

Keywords. Extreme value theory, Time series.

1 Introduction

We study stationary heavy-tailed time series with regularly varying distributions; cf. Basrak and Segers (2009). In this framework, extremal observations cluster: an extreme value

triggers a short period with numerous large observations. This behavior is known to perturb classical inference procedures tailored for independent observations like high quantile inference; see Embrechts et al. (1997). This clustering effect can be summarized with the *extremal index*, initially introduced in Leadbetter (1983) and Leadbetter et al. (1983). We can interpret it as the inverse of the mean number of consecutive exceedances above a high threshold in a short period of time. In this presentation, we aim to infer such properties of the clustering effect by letting functionals act on consecutive observations with extremal behavior. For example, we can recover the extremal index from this setting and other important indices of the extremes of the series.

2 Regularly varying time series

We consider cluster statistics of regularly varying time series $(\mathbf{X}_t)$ with values in $(\mathbb{R}^d, |\cdot|)$ and tail index $\alpha > 0$ such that $\mathbb{P}(|\mathbf{X}| > x) = \ell(x)x^{-\alpha}$, $x > 0$, and ℓ slowly varying. For cluster inference, we consider a sample $\mathbf{X}_{[1,n]}$ together with the sequence (b_n), and we define the sample of disjoint blocks $(\mathcal{B}_j)_{j=1,\dots,m_n}$ as blocks of consecutive observations with

$$\mathcal{B}_j := (\mathbf{X}_{(j-1)b_n+1}, \dots, \mathbf{X}_{jb_n}) \;=\; \mathbf{X}_{(j-1)b_n+[1\,:\,b_n]}, \tag{1}$$

such that $b_n \to \infty$, $m_n = n/b_n \to \infty$, as $n \to \infty$. We follow the clusters theory developed in Mikosch and Wintenberger (2024). The extremal behavior of regularly varying time series is modeled by the conditional behavior of a block $\mathcal{B}_j$ given that its ℓ^α-norm is large:

$$\mathbb{P}(\mathcal{B}_1/x_{b_n} \in A \,|\, \|\mathcal{B}_1\|_\alpha > x_{b_n}) \xrightarrow{w} \mathbb{P}(Y\mathbf{Q} \in A\,), \qquad n \to \infty, \tag{2}$$

such that Y is independent of $\mathbf{Q} \in \ell^\alpha$, $\mathbb{P}(Y > y) = y^{-\alpha}$, for $y > 1$, $\|\mathbf{Q}\|_\alpha = 1$ a.s. The weak convergence holds for a family of shift-invariant continuity sets $A \subset \ell^\alpha$, and (x_n) is a suitable sequence satisfying $\mathbb{P}(\|\mathcal{B}_1\|_\alpha > x_{b_n}) \to 0$, as $n \to \infty$. The spectral cluster process $\mathbf{Q}$ models the short period's behavior under the rare event that its ℓ^α-norm is large.

3 Asymptotic normality of cluster statistics

We study cluster statistics of the form

$$f_p^{\mathbf{Q}} = \mathbb{E}[f(Y\mathbf{Q})]\,, \tag{3}$$

for suitable ℓ^α-continuity functions $f : \ell^\alpha \to \mathbb{R}$, invariant to the shift operator. To infer the cluster statistic (3), we use the disjoint blocks estimators proposed in Buriticá et al. (2023), and defined as

$$\widehat{f_k^{\mathbf{Q}}} := \frac{1}{k}\sum_{t=1}^{m_n} f(\mathcal{B}_t/\|\mathcal{B}\|_{\alpha,(k+1)})\mathbf{1}(\|\mathcal{B}_t\|_\alpha > \|\mathcal{B}\|_{\alpha,(k+1)}), \tag{4}$$

where $\|\mathcal{B}\|_{\alpha,(1)} \geq \|\mathcal{B}\|_{\alpha,(2)} \geq \cdots \geq \|\mathcal{B}\|_{\alpha,(m_n)}$, denotes the sequence of order statistics of the ℓ^p-norms of blocks defined in (1).

In this presentation we establish the asymptotic normality of the block estimators in Equation (4), tailored for cluster inference. We state moment, mixing and bias assumptions yielding the existence of a sequence (k_n), satisfying $k = k_n \to \infty$, $m_n/k_n \to \infty$ such that

$$\sqrt{k}\big(\widehat{f_k^{\mathbf{Q}}} - f^{\mathbf{Q}}\big) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(f(Y\mathbf{Q}))), \quad n \to \infty,$$

and the limit is a centered Gaussian distribution. As a result, we see that the asymptotic variance of the blocks estimator can be computed in terms of the spectral cluster process $\mathbf{Q}$.

4 Comparison with existing works

We show that introducing $\ell^\alpha-$norm block order statistics in (4), instead of order statistics of the sample $(|\mathbf{X}_t|)$ as in Cissokho and Kulik (2021), and Drees and Neblung (2021), can lead to a better asymptotic variance for cluster inference. We give examples of variance reduction in the case of linear models with short-range dependence, for inference of classical indices. In our examples, the asymptotic variance $\mathrm{Var}(f(Y\mathbf{Q}))$ is null because of the deterministic properties of the spectral cluster process of linear models. For linear models, the advantage of replacing thresholds with block maxima records was previously investigated in Hsing (1993). To prove the asymptotic normality of block estimators, we rely on the asymptotics of disjoint block estimators studied in Theorem 2.10. in Drees and Rootzén (2010), and the theory of empirical processes therein. We also follow the modern overview in Kulik and Soulier (2020). To handle the asymptotics of extremal ℓ^α-blocks, we build on the large deviation principles studied in Buriticá et al. (2023), and appeal to the spectral cluster processes theory therein.

The blocks estimator in Equation (4) compares favourably to state-of-the-art methodologies for cluster inference in terms of asymptotic variance. We illustrate on simulations that the ℓ^α–block estimator is competitive both in terms of bias and variance for finite sample sizes. This approach performs well even when the tail index $\alpha > 0$ is unknown and has to be estimated with a Hill-type procedure.

Bibliographie

Basrak, B. and Segers, J. (2009) Regularly varying multivariate time series. *Stoch. Proc. Appl.* **119**, 1055–1080.

Buriticá, G., Mikosch, T., and Wintenberger, O. (2023) Large deviations of p-blocks of regularly varying time series and applications to cluster inference. *Stoch. Proc. Appl.* **161**, 68–101.

Cissokho, Y. and Kulik, R. (2021) Estimation of cluster functionals for regularly varying time series: sliding blocks estimators. *Electronic Journal of Statistics.* **15**, 2777–2831.

Drees, H. and Neblung, S. (2021) Asymptotics for sliding blocks estimators of rare events. *Bernoulli* **27**, 1239–1269.

Drees, H. and Rootzén, H. (2010) Limit theorems for empirical processes of cluster functionals. *Ann. Stat.* **38**, 2145–2186.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling extremal events for insurance and finance.* Springer-Verlab, Berlin Heidelberg New York

Hsing, T. (1993) On some estimates based on sample behavior near high level excursions. *Probab. Th. Rel. Fields.* **95**, 331–356.

Kulik, R. and Soulier, P. (2020) *Heavy-Tailed Time Series.* Springer, New York.

Leadbetter, M.R. (1983) Extremes and local dependence in stationary sequences. *Probab. Th. Relat. Fields* **65**, 291–306.

Leadbetter, M.R., Lindgren, G., and Rootzén, H. (1983) *Extremes and related properties of random sequences and processes.* Springer, Berlin.

Mikosch, T., and Wintenberger, O. (2024) *Extreme Value Theory for Time Series. Models with Power-Law Tails.* Springer, New York.

Prix ENSAI-SFdS et Etudes de cas

LIENS CAUSAUX INTERVENANT DANS LE PROCESSUS D'ADOPTION D'UN EXOSQUELETTE DANS LE CONTEXTE INDUSTRIEL

Isabelle Clerc-Urmès[1], Liên Wioland[2], Jean-Jacques Atain-Kouadio[3] & Benjamin Paty[4]

[1] *Département Sciences Appliquées au Travail et aux Organisations - Institut National de Recherche et de Sécurité, France, isabelle.urmes@inrs.fr*
[2] *Département Sciences Appliquées au Travail et aux Organisations - Institut National de Recherche et de Sécurité, France, lien.wioland@inrs.fr*
[3] *Département Sciences Appliquées au Travail et aux Organisations - Institut National de Recherche et de Sécurité, France, jean-jacques.atain-kouadio@inrs.fr*
[4] *Département Sciences Appliquées au Travail et aux Organisations - Institut National de Recherche et de Sécurité, France, benjamin.paty@inrs.fr*

Résumé.

Depuis plus de 20 ans, les troubles musculosquelettiques (TMS) représentent les maladies professionnelles les plus fréquentes en Europe touchant environ 60% des salariés (Roquelaure, Bodin et al. 2018). Dans un contexte de prévention des risques professionnels et plus spécifiquement des risques biomécaniques liés à l'activité professionnelle, lorsque les mesures de prévention collectives ne sont pas suffisantes ou ne peuvent être mises en place, des mesures individuelles telles que les dispositifs d'assistance physique et en particulier les exosquelettes peuvent alors être envisagées. Un exosquelette peut être défini comme une structure mécanique externe revêtue qui augmente la puissance ou les capacités d'une personne (de Looze, Bosch et al. 2016). Malgré un large consensus dans la littérature (Stam and Stanton 2010, Bar, Luger et al. 2024) démontrant l'efficacité des exosquelettes pour réduire localement les contraintes musculaires et le ressenti des efforts, ce n'est pas pour autant qu'en contexte professionnel les salariés l'adoptent pour un usage régulier. La question de l'adoption des exosquelettes constitue un enjeu capital dans le contexte de modernisation des entreprises où le déploiement des nouvelles technologies est devenu une question centrale pour les entreprises, d'autant plus si ces dispositifs ont un rôle à jouer dans la prévention des TMS. L'adoption d'une technologie étant le reflet de la qualité de l'interaction entre l'utilisateur et la technologie, cette étude cherche à comprendre et formaliser les mécanismes cognitifs et psychosociaux d'adoption d'un exosquelette aux étapes d'acceptabilité et d'acceptation.

Une campagne de collecte de données a été menée auprès de 181 salariés (94 non-utilisateurs de l'exosquelette et 87 utilisateurs ou anciens utilisateurs) répartis dans 15 entreprises différentes. Le questionnaire – préalablement testé auprès de 10 salariés – collecte les évaluations et le ressenti des salariés à l'égard des exosquelettes. Cet outil subjectif d'évaluation des exosquelettes permet d'identifier les éléments expliquant son adoption. Il s'articule autour de sept dimensions basées sur le modèle théorique UTAUT (Unified Theory of Acceptance and Use of Technology) (influence sociale, identité professionnelle, affects, conditions facilitantes, facilité d'utilisation perçue, performance en termes de santé et sécurité au travail, et performance en termes de production). La consistance des dimensions est vérifiée via un alpha de Cronbach. Deux modèles d'équation structurelles basés sur la méthode PLS, selon l'utilisation de l'exosquelette, déterminent les relations existantes entre l'intention d'usage de ce dernier et les différentes dimensions citées.

Il ressort de ces analyses que les mêmes dimensions sont impliquées dans les deux phases du processus d'adoption ; à savoir l'acceptabilité en amont de l'interaction du salarié avec le

dispositif et de l'acceptation en aval. Cependant, les dimensions sont organisées différemment et ont des impacts différents selon qu'il s'agit d'acceptabilité ou d'acceptation.

Les modèles ont été identifiés, leur formalisation sous forme de dimensions et leur opérationnalisation sous forme de questionnaire ont été validées. Ainsi sont mis en évidence la dynamique des différentes dimensions impliquées au cours de l'adoption des et l'identification de leviers d'action pour prévenir les risques liés à une qualité dégradée de l'interaction homme-exosquelette.

Mots-clés. Path modeling, PLS, Trouble musculosquelettique, Exosquelette, Prévention

Abstract.

For over 20 years, musculoskeletal disorders (MSDs) have been the most common occupational illnesses in Europe affecting around 60% of employees (Roquelaure, Bodin et al. 2018). In a context of preventing occupational risks, and more specifically biomechanical ones, when collective prevention measures are not sufficient or cannot be implemented, individual measures such as physical assistance devices and in particular exoskeletons can then be considered. Exoskeletons can be defined "*as a wearable, external mechanical structure that enhances the power of the user*" (De Looze et al 2016). Despite a broad consensus in the literature (Stam and Stanton 2010, Bar, Luger et al. 2024) demonstrating the effectiveness of exoskeletons in locally reducing muscular strain and the feeling of exertion, this does not mean that employees in the workplace are adopting them for regular use. The question of exoskeleton adoption is capital in the context of corporate modernization, where the deployment of new technologies has become a central issue for companies, all the more so if these devices have a role to play in MSD prevention. To adopt of a technology reflects the quality of the interaction between the user and the technology. This study seeks to understand and formalize the cognitive and psychosocial mechanisms of exoskeleton adoption at the acceptability and acceptance stages.

The sample comprised 181 employees (94 exoskeleton non-users and 87 users or former users) from 15 different companies. The questionnaire - previously tested with 10 employees - collects employees' evaluations and feelings about exoskeletons. This subjective tool evaluates exoskeletons and the factors determining their adoption. It is structured around 7 dimensions based on the UTAUT (Unified Theory of Acceptance and Use of Technology) theoretical model (social influence, professional identity, affects, facilitating conditions, perceived ease of use, performance in terms of occupational health and safety, and performance in terms of production). Internal consistency within each dimension was checked using Cronbach's alpha. Based on the PLS approach, two structural equation models, according to exoskeleton use, determine the relationships between the intention to use the exoskeleton and the various dimensions mentioned.

Same dimensions are involved in both phases of the adoption process, i.e. acceptability upstream of the employee's interaction with the exoskeleton, and acceptance downstream. However, the dimensions are organized differently and have different impacts depending on whether we're talking about acceptability or acceptance.

Models were identified, formalized in the form of dimensions and operationalized in a questionnaire. This highlighted the dynamics of the different dimensions involved during the adoption of exoskeletons, and identified levers for action to prevent the risks associated with a degraded quality of human-exoskeleton interaction.

Keywords. Path modeling, PLS, Musculoskeletal disorders, Exoskeleton, Prevention

1 Introduction

Depuis plus de 20 ans, les troubles musculosquelettiques (TMS) représentent les maladies professionnelles les plus fréquentes en Europe touchant environ 60% des salariés (Roquelaure, Bodin et al. 2018). Dans un contexte de prévention des risques professionnels et plus spécifiquement des risques biomécaniques liés à l'activité professionnelle, lorsque les mesures de prévention collectives ne sont pas suffisantes ou ne peuvent être mises en place, des mesures individuelles peuvent alors être envisagées. Ainsi, lorsque toutes les solutions d'amélioration des conditions de travail ont été testées ou que l'automatisation n'est pas envisageable, les dispositifs d'assistance physique, et plus particulièrement les exosquelettes professionnels, peuvent être envisagés comme pouvant contribuer à l'amélioration des conditions de travail. En effet, le développement des nouvelles technologies d'assistance physique et plus particulièrement des exosquelettes s'accélère.

De manière générale, un exosquelette peut être défini comme une structure mécanique externe revêtue qui augmente la puissance ou les capacités d'une personne (de Looze, Bosch et al. 2016). Il peut être qualifié *d'actif* lorsqu'il est actionné par un système robotisé, et se définit alors plus précisément comme une « structure électromécanique revêtue par l'opérateur qui s'adapte à la silhouette et aux fonctions de l'homme » (Anam and Al-Jumaily 2012). Un exosquelette est dit *passif* lorsqu'il fonctionne par restitution d'énergie mécanique grâce entre autres, à des matériaux particuliers, des systèmes à ressorts ou des élastiques (Bosch, van Eck et al. 2016). Les exosquelettes peuvent également être catégorisés en fonction de la partie du corps qu'ils assistent comme les membres supérieurs ou inférieurs, le dos, la main ou encore l'ensemble du corps (Bosch, van Eck et al. 2016). Également développés à des fins d'applications militaires et médicales (Viteckova, Kutilek et al. 2013, Yan, Cempini et al. 2015, Rupal, Rafique et al. 2017), ces systèmes émergent particulièrement dans le monde du travail ces dernières années. Ils apparaissent comme des « exosquelettes professionnels », « conçus pour apporter une assistance physique dans la réalisation d'une tâche afin de réduire leur exposition aux contraintes physiques » (Theurel, Desbrosses et al. 2018). Les entreprises les perçoivent donc comme des dispositifs pouvant contribuer à améliorer les conditions de travail, en particulier en limitant l'exposition à certains facteurs de risques biomécaniques de survenue de troubles musculosquelettiques. C'est dans ce contexte que les entreprises expérimentent ces technologies auprès des salariés. Cependant malgré un large consensus dans la littérature (Stam and Stanton 2010), démontrant l'efficacité des exosquelettes pour réduire localement les contraintes musculaires et le ressenti des efforts, les utilisateurs de ces technologies dans le contexte professionnel font part de réactions très contrastées. Une partie des salariés refusent de les tester, d'autres les utilisent puis les rejettent ; enfin, certains les intègrent dans un usage régulier. Ce constat est valable pour toutes les nouvelles technologies (NT) et il est d'ailleurs communément admis qu'environ la moitié d'entre elles ne parvient pas à trouver sa place dans le milieu professionnel (Bernoux P. 2008). Ces échecs peuvent s'expliquer par le fait que les projets de changement technologique sont encore souvent envisagés avec une représentation déterministe de la technologie qui envisage que si cette dernière est correctement développée, alors elle s'imposera de façon intrinsèque à l'organisation et aux salariés (Chevallet 2007). Ces échecs peuvent également être liés aux caractéristiques techniques de la technologie (problèmes techniques, défaillances …) ou encore au ressenti des utilisateurs potentiels. Cette étude concerne le dernier point. Son objectif est de comprendre et formaliser les mécanismes cognitifs et psychosociaux d'adoption d'un exosquelette.

Cette question constitue un enjeu capital dans le contexte de modernisation des entreprises où le déploiement des NT est devenu une question centrale pour les entreprises. Cet enjeu sera considéré à l'aune de la prévention des risques professionnels. En effet, l'introduction de toute

NT dans une organisation de travail constitue un changement majeur pour l'entreprise car elle peut modifier profondément son fonctionnement (Collerette P. 2006). Ainsi, l'introduction d'une NT peut créer de l'instabilité, susceptible de perdurer même une fois le déploiement terminé (Gollac and Volkoff 1996, Hensel R. 2019, Govaere V. 2020, Moyon A. 2020). Cette instabilité s'exprime tant au niveau des organisations (par exemple, rationalisation et formalisation de toutes activités permises par la technologie), que de l'activité des collectifs de travail, de l'encadrement et des salariés (définitions floues des rôles, des tâches et des procédures par exemple). En termes de santé et sécurité, cette instabilité peut conduire à des facteurs de dégradation tels qu'une intensification du travail, un affaiblissement des relations interpersonnelles des collectifs de travail ainsi qu'une modification des modes opératoires ou encore des cœurs de métiers (Nielsen 1993, Davezies 2006, Govaere 2013, Baltrusch, Houdijk et al. 2020, Shore, Power et al. 2020, Baltrusch, Houdijk et al. 2021). Si ces facteurs de dégradation se potentialisent, ils peuvent exposer à des facteurs de risques psychosociaux, physiques voire d'accidents. L'intégration d'un exosquelette en entreprise n'échappe pas à ce constat et mérite d'être analysée pour prévenir ces facteurs de risques. Analyser le processus d'adoption permet de traiter de ces questions. De plus, elle permet de recueillir des éléments d'information et de compréhension sur la qualité de l'interaction Homme-Exosquelette. Dans une approche de prévention des risques professionnels, la qualité de cette interaction constitue un indicateur des points de tensions (bloquants) ou d'accord (facilitateurs). Par exemple, son rejet peut révéler qu'elle n'est pas adaptée à l'activité. Il peut aussi constituer une stratégie de protection face à une activité qui se dégrade ou qui se fragilise tant au niveau individuel que collectif (danger de désorganisation, fragilisation des collectifs, risque d'isolement, perte de sens, remise en cause de l'expertise, des savoir-faire de métier, de faible communication avec l'entourage, sentiment de perte d'identité, ou encore apparition d'anxiété…). Dans certains cas, les éléments de rejet peuvent également révéler des difficultés déjà existantes dans l'organisation que l'exosquelette vient cristalliser. Il peut même devenir objet de négociation par exemple d'améliorations des conditions de travail.

Identifier les points bloquants et/ou facilitateurs permet de définir et de mettre en œuvre des pistes d'amélioration, tant au niveau de sa conception, de son intégration dans l'organisation et la situation de travail qu'au niveau de son usage et par la même, d'agir en faveur de la santé et la sécurité des utilisateurs. Outre la mise en évidence des facteurs associés à l'usage de l'exosquelette au cours de son acceptabilité ou de son acceptation, le sens de ces relations est essentiel pour l'identification des leviers d'actions pour prévenir les risques liés à une qualité dégradée de l'interaction homme-exosquelette. Dans un objectif de prévention des risques professionnels, l'objectif de cette étude est de comprendre et formaliser les mécanismes psychologiques d'adoption d'un exosquelette et en particulier les mécanismes de son acceptation ou de son rejet.

2 Matériel et méthode

Sont inclus 181 salariés (94 non-utilisateurs de l'exosquelette et 87 utilisateurs ou anciens utilisateurs). Le questionnaire – préalablement testé auprès de 10 salariés – collecte les évaluations et le ressenti des salariés à l'égard des exosquelettes, et permet d'évaluer la qualité de l'interaction Homme-Exosquelette aux étapes d'acceptabilité et d'acceptation qui composent l'adoption (Wioland L. 2019). Il s'articule autour de 7 dimensions basées sur le modèle théorique UTAUT (Unified Theory of Acceptance and Use of Technology) et celui de l'acceptation située (Bobillier-Chaumont 2013, Bobillier-Chaumon 2021) ; à savoir :

1- l'*influence sociale*, définie par le degré avec lequel un salarié perçoit l'avis des autres

concernant l'utilisation du dispositif ;
2- l'*identité professionnelle* (ou aspects professionnels et identitaires) correspondant à la représentation que les salariés se font d'eux-mêmes à partir des caractéristiques de leur travail, de leurs compétences, réalisations professionnelles, et de leur appartenance à un groupe professionnel ;
3- les *affects* reflétant les ressentis du salarié ;
4- les *conditions facilitantes* correspondant à ce que l'organisation met en place, tant en amont du déploiement des exosquelettes qu'une fois intégrés ;
5- la *facilité d'utilisation perçue* indiquant le degré selon lequel un salarié pense que l'utilisation du système ne nécessite pas d'effort et qu'il est facile d'utilisation ;
6- la *performance en termes de santé et sécurité au travail* qui évalue le degré selon lequel un salarié pense que l'utilisation d'un exosquelette améliore sa performance au travail ; et,
7- la *performance en termes de production* qui évalue le degré selon lequel un salarié pense que l'utilisation d'un exosquelette améliore sa productivité au travail.

Un α de Cronbach standardisé est calculé pour évaluer la fiabilité de la cohérence interne au sein de chacune des 7 dimensions. Les dimensions *conditions facilitantes* et *identité professionnelle* sont retirées des analyses pour les deux groupes (utilisateurs et non-utilisateurs), car l'alpha de Cronbach standardisé était inférieur à 0,7. Les autres dimensions présentent une cohérence interne acceptable comprise entre 0,709 et 0,898 (Bland and Altman 1997).

Deux modèles d'équation structurelles– selon que les salariés soient ou non utilisateurs – déterminent les relations causales existantes entre l'intention d'usage de l'exosquelette et les 5 dimensions retenues (Sobel 1982, Lohmoller 2013, Hair 2017, Venturini and Mehmetoglu 2019). Ces derniers sont basés sur l'approche PLS (Partial Least Squares). Cette approche itérative permet de modéliser des relations complexes entre variables observées et variables latentes. Les relations structurelles (via le modèle interne, i.e. entre variables latentes) et les relations de mesure (via le modèle externe, i.e. entre variables latentes et leurs indicateurs) sont simultanément estimées (Jöreskog 1982).

Le seuil de significativité des analyses est fixé à 5%. Les analyses sont effectuées à l'aide du package PLSSEM de Ventuini et al. (2019) sous le logiciel Stata version 18 (StataCorp, USA).

3 Résultats

A l'étape de l'acceptabilité, sur l'échantillon des 94 non-utilisateurs de l'exosquelette, les *affects* ($\beta = 0,359$; $p < 0,001$), l'*influence sociale* ($\beta = 0,296$; $p < 0,001$) et la *facilité d'utilisation perçue* ($\beta = 0,292$; $p < 0,01$) déterminent directement l'intention d'usage de l'exosquelette (Figure 1). Ainsi, l'intention des non-utilisateurs d'utiliser le système est principalement déterminée par la façon dont ils se représentent l'exosquelette, la réaction de leur environnement social professionnel et la facilité d'utilisation perçue de ce dernier. Plus cette représentation est positive, plus l'intention d'usage de l'exosquelette est grande. Les dimensions *affects* et *facilité d'utilisation perçue* sont particulièrement importantes à considérer. En effet, outre les liens directs qu'elles ont sur l'intention de porter le dispositif, elles s'influencent mutuellement de manière significative et positive ($\beta = 0,434$; $p < 0,001$ et $\beta = 0,422$; $p < 0,001$). Le même résultat est observé pour la dimension de l'*influence sociale* qui impacte les *affects* et vice-versa, mais dans une moindre mesure ($\beta = 0,190$; $p < 0,05$ et $\beta = 0,394$; $p < 0,05$). En outre, les *affects* et la *performance en matière de santé et de sécurité au travail* s'influencent positivement ($\beta = 0,357$; $p < 0,01$ et $\beta = 0,271$; $p < 0,01$), de même que

la *performance en matière de productivité* et la *facilité d'utilisation* (β = 0,307 ; p <0,001 et β = 0,445 ; p <0,001) (Figure 1).

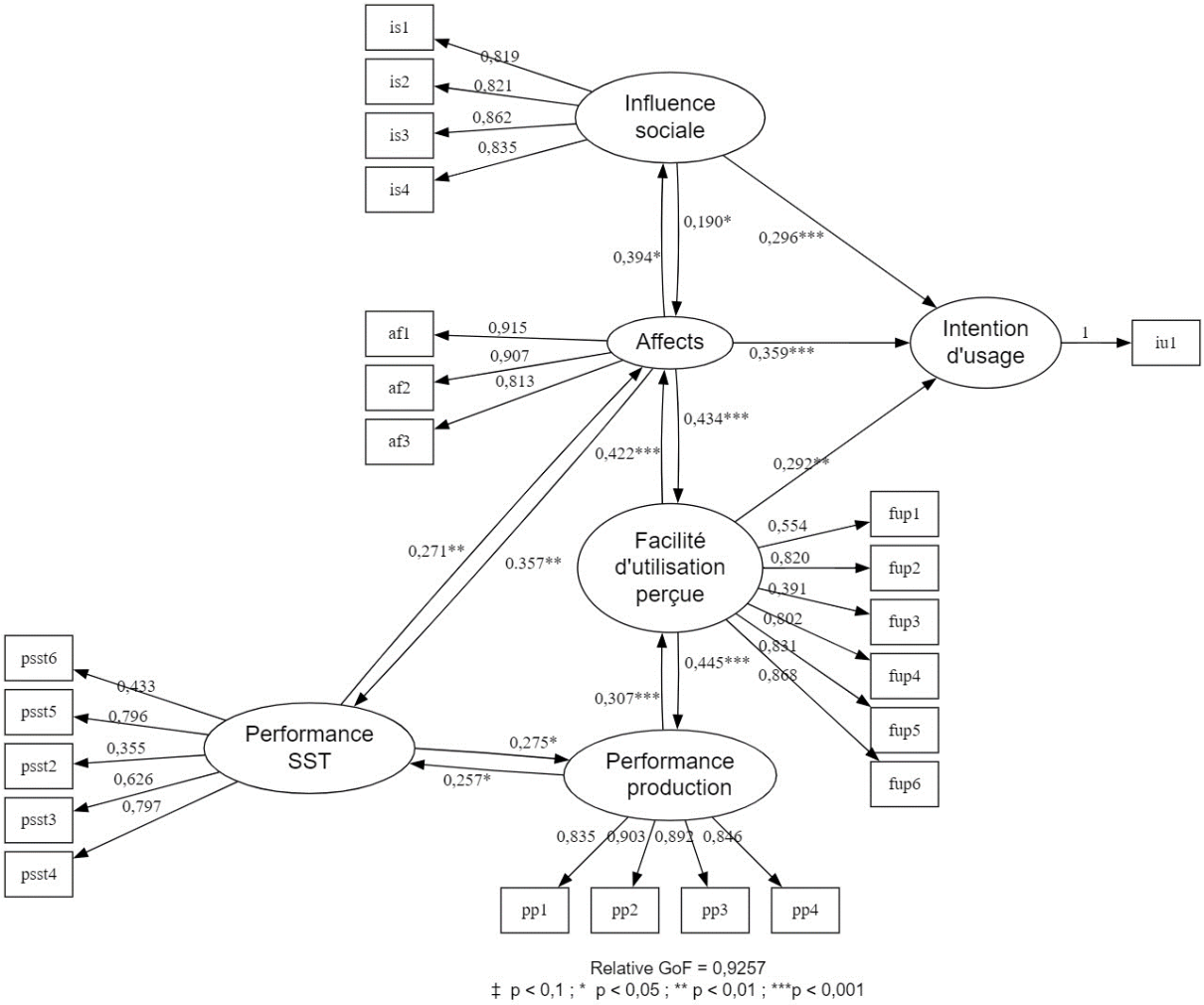


Figure 1 : Modèle d'équations structurelles des non-utilisateurs de l'exosquelette (R^2 = 0,571) (L'influence sociale est notée *is* ; les affects sont notés *af*, la facilité d'utilisation perçue est notée *fup*, les performances de production sont notées *pp*, les performances en matière de santé et de sécurité au travail sont notées *psst* et Intention d'utilisation notée *iu*).

A l'étape de l'acceptation, dans l'échantillon des 87 utilisateurs actuels et anciens, les *affects* sont la seule dimension à avoir un lien direct sur l'intention d'usage de l'exosquelette (β = 0,529 ; p <0,001) (Figure 2). Les utilisateurs ayant des affects positifs à l'égard des exosquelettes auront davantage l'intention de continuer à les utiliser dans le cadre de leur travail. Similairement aux non-utilisateurs, la *facilité d'utilisation perçue* et les *affects* s'influencent mutuellement (β = 0,248 ; p < 0,10, β = 0,261 ; p < 0,10). Contrairement aux non-utilisateurs, l'*influence sociale* n'impacte plus les *affects*. Cependant, les *affects* influencent la *performance en matière de santé et de sécurité* (β = 0,634 ; p <0,001) et inversement (β = 0,527 ; p <0,001). La *facilité d'utilisation perçue* a un lien positif réciproque avec la *productivité* (β = 0,391 ; p <0,001 et β = 0,707 ; p <0,001) et la dimension *influence sociale* (β = 0,726 ; p <0,01 et β = 0,380 ; p <0,001). Il convient de noter que *la performance en matière de productivité* et l'*influence sociale* sont également liées (β = -0,459 ; p <0,001 et β = -0,484 ; p <0,001) (Figure 2).

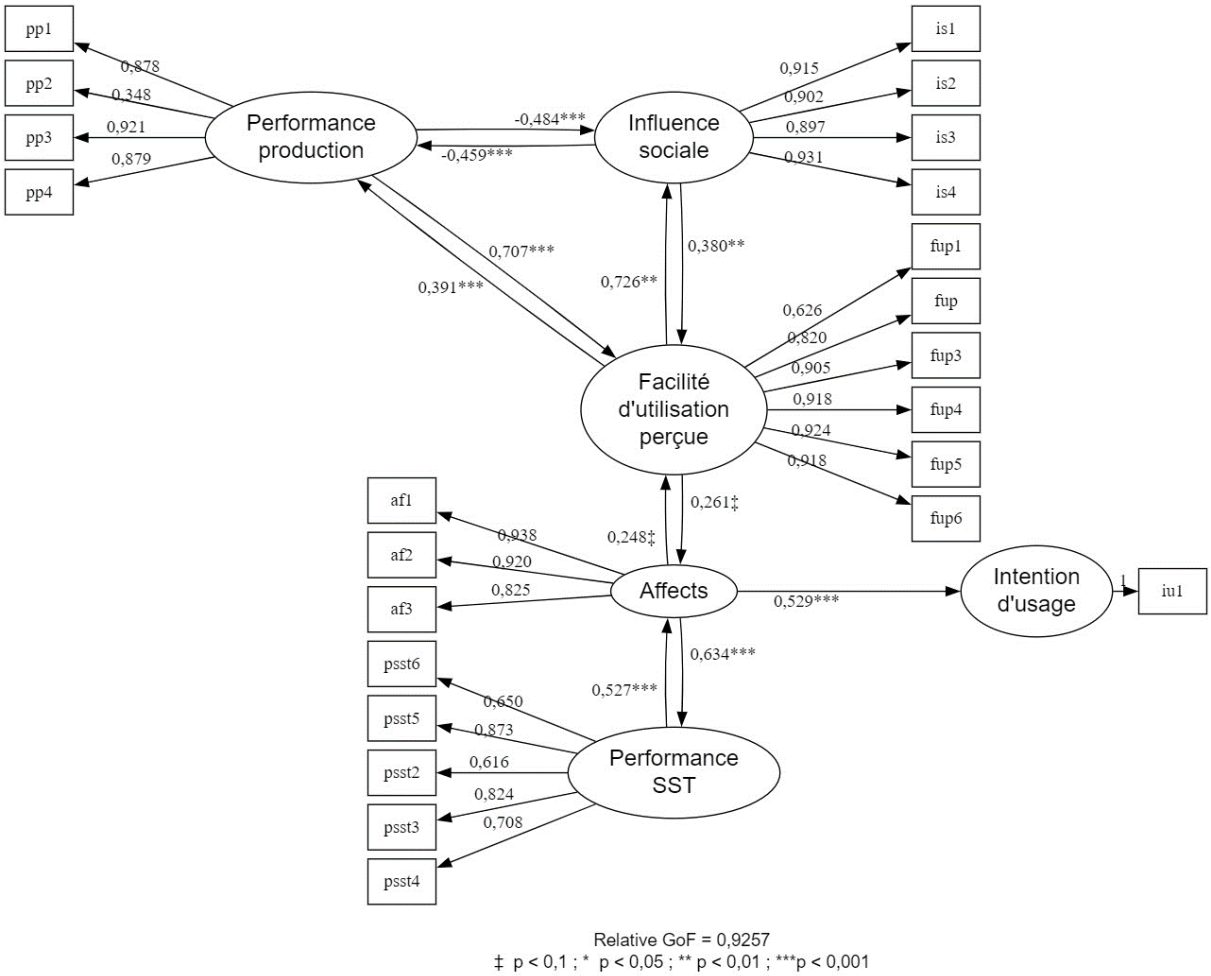


Figure 2 : Modèle d'équations structurelles des utilisateurs et anciens utilisateurs de l'exosquelette ($R^2 = 0,656$)
(L'influence sociale est notée *is* ; les affects sont notés *af*, la facilité d'utilisation perçue est notée *fup*, les performances de production sont notées *pp*, les performances en matière de santé et de sécurité au travail sont notées *psst* et Intention d'utilisation notée *iu*).

4 Conclusion

Quelle que soit l'étape de l'adoption de la technologie à laquelle une recherche s'intéresse, il est important de garder à l'esprit que le processus d'adoption d'une technologie s'inscrit dans un continuum partant de la phase d'acceptabilité en amont jusqu'à son acceptation effective et durable (continuité d'usage). L'adoption se révèle donc être un phénomène progressif et dynamique.

Malgré une tendance à sous-estimer les coefficients ce qui réduirait la puissance à détecter des relations significatives entre variables latentes, l'approche PLS des modèles d'équations structurelles permet de mettre en évidence le caractère prédictif des relations établies. Les dimensions déterminant l'acceptabilité et l'acceptation interagissent. En effet, cette étude met en évidence que l'acceptabilité et l'acceptation sont déterminées par les mêmes dimensions. En revanche, la façon dont ces dernières s'organisent, ainsi que leur poids sur l'intention d'usage

diffèrent selon l'étape du processus d'adoption à laquelle se trouve l'utilisateur. Ce processus apparaît donc comme dynamique. Toutefois cette étude ne permet pas d'appréhender les mécanismes qui sous-tendent cette dynamique. Pour se faire un suivi longitudinal aux différentes étapes du processus d'adoption des exosquelettes des salariés devrait être envisagé. Le croisement des séries chronologiques et des modèles d'équation structurelles pourrait être une méthode à développer pour caractériser simultanément l'intention d'usage à différents temps du processus dynamique d'adoption de la technologie.

Intégrer un exosquelette en entreprise implique de considérer la question de l'adoption du dispositif par les salariés mais il est également important d'appréhender les questions relatives aux dynamiques collectives susceptibles d'influencer le choix de l'utilisateur à adopter ou non le dispositif, ainsi la manière dont l'encadrement intègrent les modifications de modes opératoires et les stratégies individuelles et collectives dans l'organisation.

5 Bibliographie

Anam, K. and A. A. Al-Jumaily (2012). "Active Exoskeleton Control Systems: State of the Art." *International Symposium on Robotics and Intelligent Sensors 2012 (Iris 2012)* **41**: 988-994.

Baltrusch, S. J., H. Houdijk, J. H. van Dieën and J. T. C. M. de Kruif (2021). "Passive Trunk Exoskeleton Acceptability and Effects on Self-efficacy in Employees with Low-Back Pain: A Mixed Method Approach." *Journal of Occupational Rehabilitation* **31**(1): 129-141.

Baltrusch, S. J., H. Houdijk, J. H. van Dieen, C. A. M. van Bennekom and A. de Kruif (2020). "Perspectives of End Users on the Potential Use of Trunk Exoskeletons for People With Low-Back Pain: A Focus Group Study." *Hum Factors* **62**(3): 365-376.

Bar, M., T. Luger, R. Seibt, M. A. Rieger and B. Steinhilber (2024). "Using a Passive Back Exoskeleton During a Simulated Sorting Task: Influence on Muscle Activity, Posture, and Heart Rate." *Hum Factors* **66**(1): 40-55.

Bernoux P., G. Y.-C. (2008). "Une nouvelle voie pour réussir les changements technologiques : la co-construction." *La Revue des Sciences de Gestion* **232**(5): 51-58.

Bland, J. M. and D. G. Altman (1997). "Cronbach's alpha." *British Medical Journal* **314**(7080): 572-572.

Bobillier-Chaumon, M. E. (2021). Exploring the situated acceptance of emerging technologies in and concerning activity: Approaches and processes. In Digital transformations in the challenge of activity and work: Understanding and supporting technological changes.

Bobillier-Chaumont, M. E. (2013). Condition d'usage et facteurs acceptation des technologies dans l'activité : questions et perspectives pour la psychologie du travail. HDR, Université Pierre Mendès-France, Grenoble, France.

Bosch, T., J. van Eck, K. Knitel and M. de Looze (2016). "The effects of a passive exoskeleton on muscle activity, discomfort and endurance time in forward bending work." *Applied Ergonomics* **54**: 212-217.

Chevallet, R. (2007). "Réussir un projet système d'information en PME : l'enjeu des conditions de travail." *ANACT*.

Collerette P., L. P., Manghi M. (2006). "A successful IT change in a police service." *Journal of Change Management* **6**(2): 159-179.

Davezies, P. (2006). "Les coûts de l'intensification du travail." *Santé et Travail* **57**.

de Looze, M. P., T. Bosch, F. Krause, K. S. Stadler and L. W. O'Sullivan (2016). "Exoskeletons for industrial application and their potential effects on physical work load." *Ergonomics* **59**(5): 671-681.

Gollac, M. and S. Volkoff (1996). "Citius, altius, fortius." *Actes De La Recherche En Sciences Sociales*(114): 54-&.

Govaere, V. (2013). "Changements de l'activité de travail liés aux NTIC : intérêts et limites du phéomène de propagation pour la prévention." *Références en santé au travail* **133**: 43-58.

Govaere V., Wioland. L. (2020). "Etude d'intégration d'un système informatique embarqué (SIE) dans le transport routier de marchandises." *Revue Hygiène et Sécurité au Travail* **260** (septembre 2020).

Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M. (2017). A primer on partial least squares structural equation modeling (PLS-SEM).

Hensel R., K. M. (2019). "Subjective evaluation of a passive industrial exoskeleton for lower-back support: a filed study in the automotive sector." *IISE transactions on occupational ergonomics and human factors* **7**(3-4): 213-221.

Jöreskog, K.G. and Wold, H. (1982). "The ML and PLS Techniques for Modeling with Latent Variables: Historical and Comparative Aspects." In: K.G. Jöreskog and H. Wold (Eds.), "Systems Under Indirect Observation", Part 1, North-Holland, Amsterdam, 263-270.

Lohmoller, J. B. (2013). Latent variable path modeling with partial least squares. , Springer Science & Business Media.

Moyon A., P. J. F., Poirson E. (2020). "Investigating the effects of passive exoskeletons and familiarization protocols on arm-elevated tasks." *Proceeding of the human factors and ergonomics society Europe*: 1-17.

Nielsen, J. (1993). Usability engineering.Boston: Academic Press.

Roquelaure, Y., J. Bodin, A. Descatha and A. Petit (2018). "[Musculoskeletal disorders: how to recognize them as occupational diseas]." *Rev Prat* **68**(10): 1132-1134.

Rupal, B. S., S. Rafique, A. Singla, E. Singla, M. Isaksson and G. S. Virk (2017). "Lower-limb exoskeletons: Research trends and regulatory guidelines in medical and non-medical applications." *International Journal of Advanced Robotic Systems* **14**(6).

Shore, L., V. Power, B. Hartigan, S. Schulein, E. Graf, A. de Eyto and L. O'Sullivan (2020). "Exoscore: A Design Tool to Evaluate Factors Associated With Technology Acceptance of Soft Lower Limb Exosuits by Older Adults." *Hum Factors* **62**(3): 391-410.

Sobel, M. N. (1982). Asymptotic confidence intervals for indirect effects in structural equations models. *Sociological methodology*, Jossey-Bass**:** 290-312.

Stam, K. R. and J. M. Stanton (2010). "Events, emotions, and technology: examining acceptance of workplace technology changes." *Information Technology & People* **23**(1): 23-53.

Theurel, J., K. Desbrosses, T. Roux and A. Savescu (2018). "Physiological consequences of using an upper limb exoskeleton during manual handling tasks." *Applied Ergonomics* **67**: 211-217.

Venturini, S. and M. Mehmetoglu (2019). "plssem: A Stata Package for Structural Equation Modeling with Partial Least Squares." *Journal of Statistical Software* **88**(8): 1-35.

Viteckova, S., P. Kutilek and M. Jirina (2013). "Wearable lower limb robotics: A review." *Biocybernetics and Biomedical Engineering* **33**(2): 96-105.

Wioland L., D. L., Atain-Kouadio J.J. (2019). "Processus d'acceptabilité et d'acceptation des exosquelettes : évaluation par questionnaire." *Références en santé au travail* **160**: 49-76.

Yan, T. F., M. Cempini, C. M. Oddo and N. Vitiello (2015). "Review of assistive strategies in powered lower-limb orthoses and exoskeletons." *Robotics and Autonomous Systems* **64**: 120-136.

Identification du réseau de corrélations des variables potentiellement influentes pour les biosolutions, premières modélisations

Pauline Dusfour-Castan [1,3] & Bénédicte Fontez [2,3] & Elsa Ballini [4] & Patrice Loisel [3,5]

[1] *pauline.dusfour-castan@supagro.fr*
[2] *benedicte.fontez@supagro.fr*
[3] *MISTEA, INRAE, Institut Agro, Univ Montpellier, 2 Place Viala 34060 Montpellier*
[4] *PHIM, CIRAD, Institut Agro, Univ Montpellier, Chem. de Baillarguet, 34980 Montferrier-sur-Lez, elsa.ballini@supagro.fr*
[5] *patrice.loisel@inrae.fr*

Résumé. Afin de lutter contre les pathogènes tout en respectant l'environnement, il est essentiel de développer des stratégies basées sur l'immunité des plantes. De nombreuses biosolutions (traitements agissant sur la santé des plantes ou sur sa nutrition) ont été testées en laboratoire, offrant de nouvelles perspectives pour l'agriculture. Cependant, ces essais simplifiés ne tiennent pas compte des interactions complexes entre la plante et son environnement, entraînant des résultats souvent décevants sur le terrain. L'approche proposée vise pour l'instant à d'identifier et caractériser les facteurs et variables potentiellement influentes afin de mieux comprendre leurs interactions. L'hétérogénéité des jeux de données issues de diverses expérimentations pose un défi majeur en matière d'harmonisation et d'analyse. Deux approches complémentaires ont été utilisées pour analyser respectivement les facteurs environnementaux/expérimentaux et l'état physiologique de la plante. La première repose sur l'Analyse en Composantes Principales Bayésienne, une extension de l'ACP classique intégrant des principes bayésiens pour améliorer la robustesse des analyses face aux données bruitées et incomplètes. La seconde approche concerne l'inférence de réseaux de co-expression génique, permettant d'explorer les relations entre les gènes en se basant sur des mesures de corrélation. En perspective, l'objectif sera de proposer une approche globale combinant ces deux types d'analyse afin d'extraire de la connaissance et mieux comprendre les interactions entre les conditions environnementales, l'état physiologique des plantes et l'expression génique afin d'optimiser l'utilisation des biosolutions dans un contexte agricole réel. Des premières modélisations de réseaux vous seront notamment présentées.

Mots-clés. Agro-écologie, ACP Bayésienne, Réseau de gènes, Biosolutions

Abstract. To combat pathogens while respecting the environment, it is essential to develop strategies based on plant immunity. Numerous biosolutions (treatments acting on plant health or nutrition) have been tested in laboratories, offering new prospects for agriculture. However, these simplified trials do not take into account the complex interactions between the plant and its environment, often leading to disappointing results in the field. The proposed approach aims to identify and characterize potentially influential factors and variables to better understand their interactions. The heterogeneity of datasets from various experiments presents a major challenge in terms of harmonization and analysis. Two complementary approaches have been used to analyze, respectively, environmental/experimental

factors and the physiological state of the plant. The first approach is based on Bayesian Principal Component Analysis, an extension of classical PCA that integrates Bayesian principles to improve the robustness of analyses in the presence of noisy and incomplete data. The second approach focuses on gene co-expression network inference, allowing the exploration of relationships between genes based on correlation measurements. Looking ahead, the goal is to propose a comprehensive approach combining these two types of analysis to extract knowledge and better understand the interactions between environmental conditions, plant physiological state, and gene expression, ultimately optimizing the use of biosolutions in real agricultural settings. Initial modeling of networks will also be presented.

Keywords. Agro-ecology, Bayesian PCA, Gene network, Biosolutions

1 Contexte et objectifs

De nombreuses biosolutions ont été testées en laboratoire mettant en lumière de nouvelles perspectives de pratiques agricoles. Toutefois, ces essais, réalisés dans des conditions d'expérimentations simplifiées, ne prennent pas en compte les multiples et complexes interactions existantes entre la plante et son evironnement. Les résultats obtenus sur le terrain sont alors parfois décevants Certaines catégories de facteurs d'efficacité ont toutefois étaient identifiées : (i) les conditions d'applications, (ii) les conditions environnementales, (iii) l'état physiologique de la plante. L'objectif de cette étude, qui n'est qu'à ses débuts, est alors d'identifier, de caractériser et de hiérarchiser ces facteurs afin de mieux comprendre les intéractions qui les lient. Pour cela, nous proposons d'étudier des variables relatives aux conditions environnementales, au rendement, à l'état physiologique de la plante, ainsi qu'à son transcriptome. Cependant, les jeux de données disponibles pour ces analyses sont souvent complexes et hétérogènes, ne contenant pas systématiquement les mêmes types de variables ce qui soulève l'enjeu de leur harmonisation et de l'élaboration de méthodes adaptées pour en tirer des conclusions robustes.

2 Matériel et données

Trois jeux de données issus d'expérimentation sur le blé ont été mis à notre disposition. Le jeu de données "Récolte 2022" concerne l'effet de deux biosolutions évaluées sur trois variétés de blé cultivées soit en monoculture soit en mélange. Ce jeu de données comprend 96 lignes, réprésentant des chaumes de blé, et 21 variables relatives à la maladie, au rendement et au sol. Le jeu de données "Récolte 2023" concerne deux biosolutions évaluées sur cinq variétés de blé. Ce jeu de données est composé de 300 lignes et de 15 variables relatives à la maladie, au rendement, et au sol. Enfin, le jeu de données "Expression génique 2022" porte sur les niveaux d'expression de différents gènes. Il se compose de 89 colonnes représentant différents gènes et de 85 lignes correspondant à des chaumes de blé.

3 Modèles

L'objectif de l'approche bayeisenne proposée est d'identifier les correlations importantes entre les variables quantitatives d'un système agronomique complexe. L'expertise sur ce type de variables issues d'expérimentation au champs n'est pas suffisante et trop parcellaire pour construire un prior ou même un modèle causal. Deux approches ont été implémentées. La première consiste en une ACP Bayésienne avec contrainte de parcimonie afin d'obtenir les densités a posteriori des corrélations entre variables à partir des premières composantes, afin d'établir un réseau de corrélations. Cette approche peut s'appliquer aux deux types de données (récolte et génétique). Pour les données génétiques, d'autres approches d'inférence de réseaux ont été étudiées afin d'apporter d'autres types d'informations (liens non linéaires par exemple). Ces réseaux peuvent servir de comparaison avec le réseau obtenu par ACP.

3.1 ACP Bayésienne

L'Analyse en Composantes Principales Bayésienne (Bayesian PCA), introduite par Christopher M. Bishop (1998), est une extension de l'ACP classique qui intègre des principes bayésiens afin d'améliorer l'analyse des données.

A partir d'un ensemble de données X de dimension $N \times D$, cette approche consiste à introduire une matrice Z de dimension $N \times q$ tel que

$$X_i = Z_i \times W' + \epsilon_i$$

avec q le nombre de composantes principales sélectionnées, W une matrice de taille $D \times q$, et $\epsilon \sim \mathcal{N}(0, \sigma^2 I_D)$ un bruit blanc gaussien.

Les priors utilisés sont les suivants : $Z_i \sim \mathcal{N}(0, I_q)$; $\sigma \sim \mathcal{IG}(0.001, 0.001)$;
$W_j|\lambda_j,\ \tau,\ c \sim \mathcal{N}(0, \tau^2\tilde{\lambda}_j^2)$; $\tau \sim \mathcal{C}^+(0, 0.5)$; $\lambda_j \sim \mathcal{C}^+(0, 1)$; $c \sim \mathcal{IG}(2, 5)$,
et $\tilde{\lambda}_j^2 = \dfrac{c^2\lambda_j^2}{c^2 + \tau^2\lambda_j^2}$.

Un prior regularized horseshoe (Piironen, J. & Vehtari, A. (2017)) a été testé pour ajouter des contraintes de parcimonie à la fois globalement (sélection automatique des axes) et localement (contribution des variables aux axes). L'utilisation de ce type de prior a été testée et évaluée par simulations.
Un des atouts majeurs de cette approche réside dans l'accès aux densités de probabilité des corrélations entre variables. Cela permet d'évaluer la probabilité qu'une corrélation donnée se situe dans un certain intervalle et d'identifier des relations structurelles entre les variables.

Cette approche a été implémentée sur le jeu de données "Récolte 2022" avec un nombre de composantes retenues égal à 4. Ci-dessous, des exemples de densités de probabilité des corrélations entre la variable "Yield" et les variables "Proteins", "Weight", et "Pest_Incidence_3".

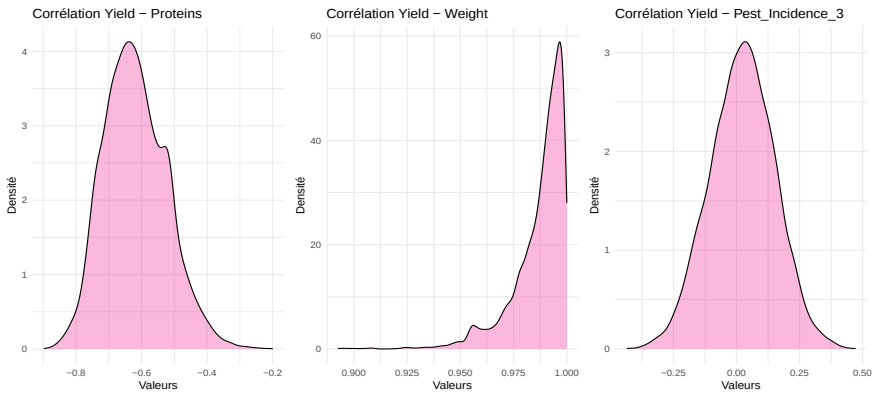


Figure 1: Densité de probabilité des corrélations entre la variable Yield et les variables Proteins, Weight, et Pest_Incidence_3, obtenues à l'aide de l'ACP Bayésienne

La figure 5 illustre un exemple de graphe de variables construit à partir de l'analyse des corrélations obtenues. Une arête a été placée entre deux variables lorsque la probabilité que leur corrélation appartienne à l'intervalle $[-1, -0.7]$, ou $[0.7, 1]$ était supérieure à 0.5. Les blocs de variables ont été identifiés à l'aide d'un modèle stochastique de blocs (SBM), qui permet de détecter automatiquement des blocs de variables fortement interconnectées. Sur le plan biologique, ces blocs correspondent à des sous-ensembles cohérents de variables (par exemple, critères agronomiques ou indicateurs de maladies), et révèlent également une organisation temporelle implicite, bien que l'information sur les dates de mesure n'ait pas été fournie au modèle.

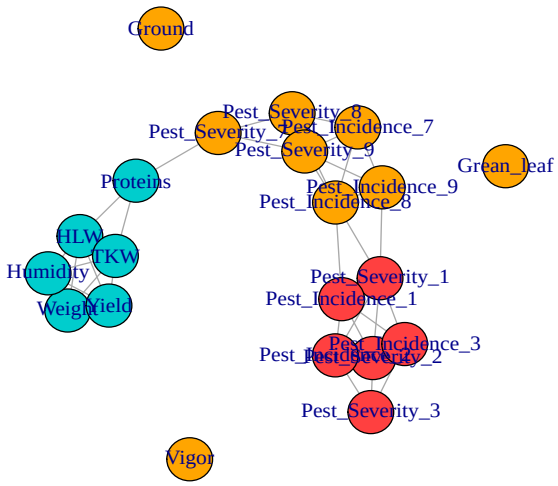
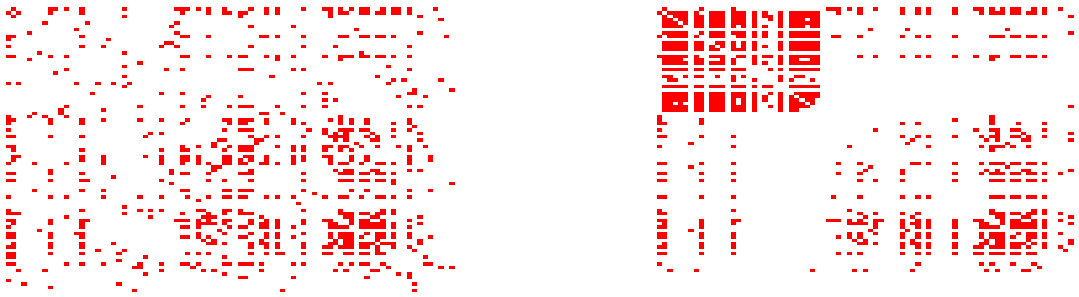


Figure 2: Graphe de corrélations, issues de l'ACP Bayésienne, entre variables et identification de blocs (SBM)

3.2 Réseaux de co-expression génique

Pour analyser les données transcriptomiques, des méthodes d'inférence de réseaux de co-expression génique ont tout d'abord été utilisées. Plus précisément, nous nous sommes appuyés sur deux types de réseaux : ceux construits à partir des corrélations de Pearson entre gènes et ceux basés sur les corrélations partielles.
Ces approches donnent lieu à des graphes où les liens peuvent être directs ou indirects, linéaires ou non linéaires. Trois méthodes ont été appliquées, une par régression bivariée scaled Lasso qui a été proposée par Ren et al (2015), une basée sur les arbres de régression proposée par Huynh-Thu et al (2010), et une basée sur les corrélations de Pearson. Pour

implémenter ces approches, les packages `SILGGM 1.0.0` (Zhang et al (2018)) et `GENIE3 1.26.0` (Huynh-Thu et al (2010)) ont été utilisés.
De la même manière que pour les données "Récolte 2022", une ACP Bayésienne a été réalisée afin d'inférer un réseau de corrélations.
Pour visualiser les différences structurelles entre les réseaux obtenus, des heatmaps de la somme des matrices d'adjacence ont été réalisées.



(a) Corrélations partielles VS corrélations issues de l'ACP

(b) Corrélations de Pearson VS corrélations issues de l'ACP

Figure 3: Comparaison sous forme de heatmaps des réseaux de gènes obtenus. Les cases rouges indiquent les arêtes présentes dans un seul des deux réseaux. Les cases blanches correspondent aux situations où les deux réseaux partagent le même état, que ce soit la présence ou l'absence d'une arête.

4 Perspectives

Bien que ce projet en soit encore à ses premiers stades, l'intégration conjointe des données environnementales, physiologiques et transcriptomiques constitue un défi méthodologique de taille. Une perspective en cours concerne le couplage des tableaux de variables expérimentales et des tableaux de gènes. Dans notre exemple, pour 2022, sur les 85 individus présents à la fois dans les tableaux "Récolte 2022" et "Expression génique 2022", nous pouvons faire des analyses multitableaux. Pour ce faire, les composantes issues de chaque ACP Bayésienne sont récupérées puis couplées afin de reformer un espace latent, dans lequel les corrélations entre variables expérimentales et gènes pourront être explorées. Les résultats de cette première proposition d'analyse seront présentés et commentés.

5 Annexe

5.1 Présentation du modèle d'ACP Bayésienne

L'Analyse en Composantes Principales Bayésienne (ACP Bayésienne), introduite par Christopher M. Bishop (1998), est une extension de l'ACP classique qui intègre des principes bayésiens afin d'améliorer l'analyse des données.

A partir d'un ensemble de données X de dimension $N \times D$, cette approche consiste à introduire une matrice Z de dimension $N \times q$ tel que

$$X_i' = W \times Z_i' + \epsilon_i$$

avec q le nombre de composantes principales sélectionnées, W une matrice de taille $D \times q$, et $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I_D)$ un bruit blanc gaussien.

Des priors sont posés sur les paramètres inconnus :

- $Z_i \sim \mathcal{N}(0, I_q)$;
- $\sigma \sim \mathcal{IG}(a_\sigma, b_\sigma)$;
- $W_j | \lambda_j,\ \tau,\ c \sim \mathcal{N}(0, \tau^2 \tilde{\lambda}_j^2)$;
- $\tau \sim \mathcal{C}^+(0, b_\tau)$;
- $\lambda_j \sim \mathcal{C}^+(0, 1)$;
- $c \sim \mathcal{IG}(a_c, b_c)$

avec $\tilde{\lambda}_j^2 = \dfrac{c^2 \lambda_j^2}{c^2 + \tau^2 \lambda_j^2}$.

Nous avons utilisé un prior regularized horseshoe (Piironen, J. & Vehtari, A. (2017)) sur les coefficients de la matrice des loadings W, afin de favoriser une représentation parcimonieuse tout en évitant des valeurs extrêmes irréalistes. Ce prior est conçu pour favoriser les coefficients proches de zéro tout en permettant à certains d'entre eux de s'en éloigner significativement. La contrainte de parcimonie globale permet d'identifier un petit nombre de dimensions latentes réellement informatives, tandis que la régularisation locale évite de surestimer la contribution de certaines variables aux axes.

5.2 Premiers résultats d'une simulation

Afin d'évaluer les performances de l'ACP bayésienne dans un cadre contrôlé, nous avons simulé des jeux de données à partir d'une structure latente de faible dimension, selon le modèle suivant :

$$X_i = W \times Z_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2 I_D)$$

où

- $X_i \in R^D$ est le vecteur d'observations simulés pour l'individu i
- $Z_i \in R^q$ est un vecteur latent simulé selon une loi normale $\sim \mathcal{N}(0, I_q)$;
- $W \in R^{D \times q}$ est la matrice des loadings, avec des coefficients tirés indépendamment selon une loi normale $\sim \mathcal{N}(0, \beta^2)$;
- ϵ_i est un bruit blanc gaussien indépendant, avec σ fixé

Les paramètres utilisés pour la simulation suivante sont (avec graine fixée):

- $N = 100$;
- $D = 30$;
- $q = 3$
- $\beta = 2$
- $\sigma = 0.5$

L'objectif est d'évaluer la capacité de l'ACP bayésienne à: (i) reconstruire fidèlement les données X, (ii) retrouver la structure latente W, (iii) estimer une matrice de corrélation entre variables qui reflète correctement la structure de celle estimée sur les données.

La figure 1 ci-dessous présente la matrice des loadings estimée par l'ACP bayésienne sous forme de heatmap, chaque case représentant le coefficient d'une variable (ligne) sur une dimension latente (colonne). Les zones colorées (rouge pour les valeurs positives, bleu pour les négatives) indiquent une contribution significative des variables aux composantes latentes. On observe une structure parcimonieuse, avec un nombre limité de composantes concentrant l'essentiel des valeurs non nulles. Trois dimensions sont clairement identifiées comme majoritaires, ce qui est en accord avec la structure latente simulée. Les autres dimensions présentent des valeurs proches de zéro, traduisant une parcimonie dans l'estimation, ce qui est attendu compte tenu de l'utilisation d'un prior regularized Horseshoe.

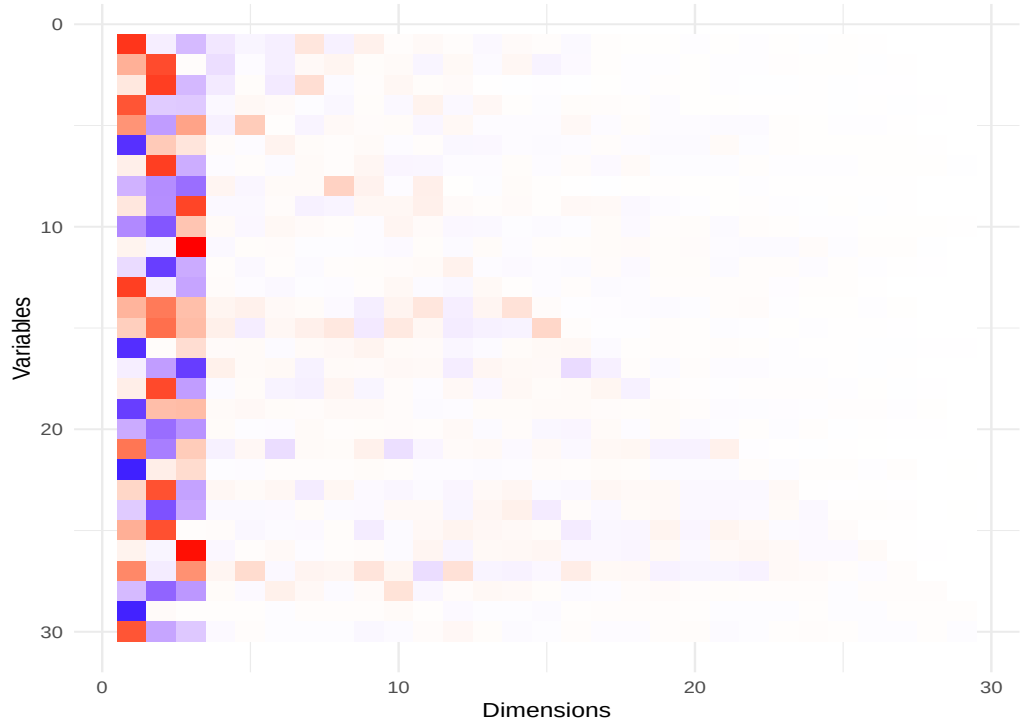


Figure 4: Structure de la matrice des loadings estimée dans le cadre de l'ACP bayésienne. Chaque case représente le coefficient de chargement d'une variable sur une dimension latente. Les valeurs proches de 0 apparaissent en blanc, tandis que les valeurs positives (rouge) et négatives (bleu) traduisent une contribution significative à la dimension considérée.

L'erreur quadratique moyenne (RMSE) entre les données simulées et reconstruites est de 0.59, ce qui est proche de la variance du bruit simulé ($\sigma = 0.5$), suggérant une reconstruction fidèle des données et une bonne récupération de la structure latente par le modèle.

En complément de l'analyse de la reconstruction des données, la figure 2 compare la matrice de corrélation du modèle (gauche) à celle estimée par l'ACP bayésienne (droite), via la matrice de chargements. Visuellement, on observe une forte similitude entre les deux structures, notamment dans la répartition des corrélations positives (rouge) et négatives (bleu), confirmant que le modèle parvient à restituer fidèlement les dépendances entre variables.

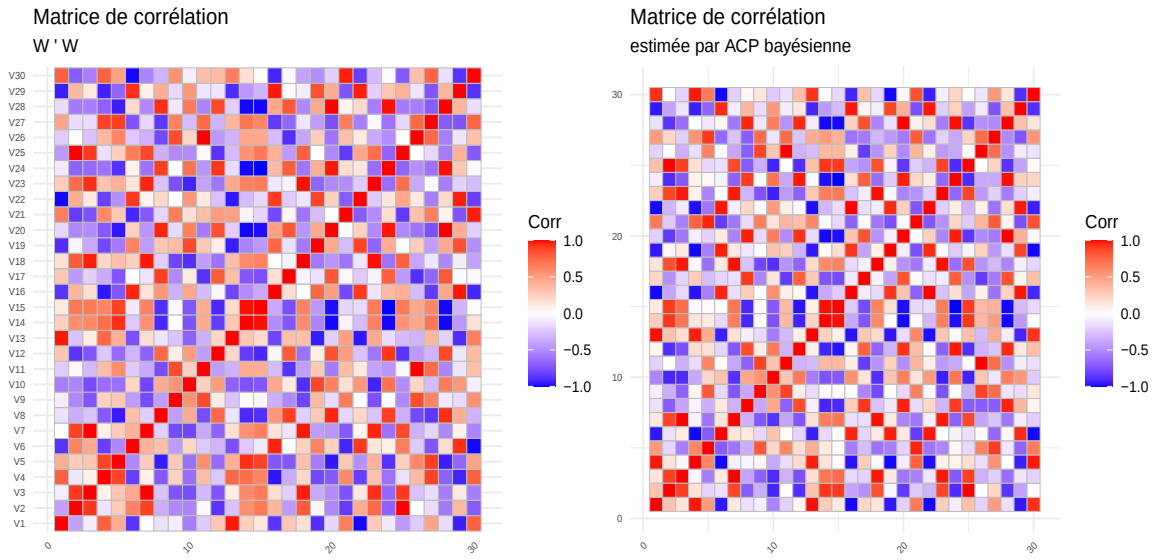


Figure 5: Heatmap des corrélations

Ce travail constitue une première exploration de l'Analyse en Composantes Principales bayésienne avec prior regularized Horseshoe. Les résultats obtenus dans un cadre simulé sont encourageants : le modèle parvient à reconstruire fidèlement les données, à retrouver les composantes latentes et à restituer la structure de corrélation entre variables. Cependant, cette approche reste encore à approfondir. Il conviendra notamment d'étudier plus finement ses propriétés selon différents régimes de bruit, tailles d'échantillons ou niveaux de parcimonie, ainsi que de mieux formaliser les critères d'évaluation, en particulier pour l'estimation des matrices de corrélation. À terme, l'objectif est d'appliquer ce cadre à des données réelles et hétérogènes issues d'expérimentations en conditions de terrain, afin de mieux comprendre les interactions entre les variables environnementales, physiologiques et géniques. Ce travail constitue ainsi une base méthodologique sur laquelle s'appuyer pour de futures analyses.

Bibliographie

Bishop C. (1998), Bayesian pca, *Advances in neural information processing systems*, 11

Ren, Z., Sun, T., Zhang, C.-H., & Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics*.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, vol. 5, no 9, p. e12776

Zhang, R., Ren, Z., & Chen, W. (2018). SILGGM: An extensive R package for efficient statistical inference in large-scale gene networks. *PLoS computational biology*, vol. 14, no 8, p. e1006369.

Piironen, J. & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors.

Bien-être des porcs en élevages alternatifs en France

Vito Toti*[1], Maxime Tio*[1], and Nathan Saraiva[1]

[1]Ecole Nationale de la Statistique et de l'Analyse de l'Information [Bruz] – Ensai, Ecole Nationale de la Statistique et de l'Analyse de l'Information – France

Résumé

Les élevages alternatifs de porcs connaissent un essor significatif en France et répondent à la demande croissante des ménages soucieux de la qualité et de l'origine des produits qu'ils consomment. Toutefois, ces systèmes ont des contraintes différentes par rapport aux élevages intensifs, et comportent des risques pour la santé animale en raison d'une exposition accrue aux maladies et aux agents pathogènes. Ces conditions mettent en péril le bien-être des porcs et peuvent causer des ravages considérables chez les animaux. C'est dans ce contexte que le programme PIGAL, au sein duquel nous nous inscrivons, a pour objectif d'étudier les facteurs qui contribuent le plus au bien-être dans ces environnements. Nous avons procédé au traitement et à l'analyse d'échantillons recoupant différents types d'élevages, à savoir les élevages de truies reproductrices (Naisseurs), les élevages de porcs en croissance (Engraisseurs) et les élevages combinant les deux activités (Naisseurs-Engraisseurs). Compte tenu du grand nombre de variables explicatives et du nombre limité d'élevages étudiés, nous avons utilisé des méthodes adaptées à ce type de données, telles que la régression PLS multiblocs et la régression Elastic Net. Nos travaux ont mis en évidence une variété kaléidoscopique de facteurs de bien-être animal, regroupés au sein de thématiques comprenant l'alimentation, l'hygiène, le logement ou encore la santé. Ces résultats serviront à élaborer des recommandations destinées aux éleveurs et aux pouvoirs publics, afin d'éclairer leurs décisions et améliorer ainsi le bien-être des porcs dans les élevages alternatifs.

Mots-Clés: Projet, Statistique, Porcs, Elevages, ENSAI, Bien, être

*Intervenant

Évaluation de modèles d'apprentissage automatique pour la prévision de la survenue d'accidents du travail à partir de séries temporelles binaires

Aho Yapi† [1,2], Pierre Latouche [1,3], Arnaud Guillin [1] & Yan Bailly [2]

[1] *Université Clermont Auvergne, CNRS, Laboratoire LMBP, UMR6620, Aubière, France*
[2] *CIKABA, Solutions QHSE, Clermont-Ferrand, France*
[3] *Institut Universitaire de France (IUF)*

Résumé. Les accidents du travail représentent un enjeu humain et économique important, mais les stratégies actuelles de prévention manquent d'anticipation proactive. Les données issues des remontées terrain, quasi-accidents et facteurs organisationnels sont sous-exploitées. Cette étude propose un cadre générique de prévision à court terme des accidents, basé sur des séries temporelles binaires et intégrant des covariables. Une validation croisée spécifique aux séries temporelles et une évaluation par périodes agrégées assurent la robustesse des modèles. Plusieurs algorithmes d'apprentissage automatique ont été comparés. Les résultats confirment la pertinence de l'approche et aident à identifier les modèles les plus efficaces pour prévenir les accidents du travail.

Mots-clés. Prévention des accidents du travail, séries temporelles binaires, apprentissage automatique

Abstract. Workplace accidents represent a significant human and economic challenge, but current prevention strategies lack proactive anticipation. Data from field reports, near-misses, and organizational factors remain underutilized. This study proposes a generic framework for short-term accident prediction, based on binary time series and incorporating covariates. A time-series-specific cross-validation methodology and evaluation using aggregated periods ensure the robustness of the models. Several machine learning algorithms were compared. The results confirm the relevance of the approach and help identify the most effective models for preventing workplace accidents.

Keywords. Workplace accident prevention, binary time series, machine learning

1 Introduction

Un accident du travail se produit lorsqu'un événement, lié à l'activité professionnelle, provoque une lésion physique ou psychologique chez un employé. En France, selon les rapport de l'assurance maladie [1], plus de 600 000 accidents du travail ont été recensés en 2023, dont près de 700 mortels, et ces chiffres ne montrent pas de tendance significative à la baisse au cours des dernières années. Ce constat montre l'insuffisance des stratégies

†. email : A-Aymar.YAPI@doctorant.uca.fr

actuelles pour réduire les accidents du travail. Depuis les travaux d'Heinrich [7], qui ont introduit l'un des premiers modèles d'analyse des accidents dénomée la théorie des dominos, la perception des accidents a évolué. Ceux-ci ne sont plus uniquement considérés comme des événements imprévisibles, mais plutôt comme la conséquence d'une série de facteurs avec des intéractions souvent complexes. Par là, de nombreux auteurs ont cherché à modéliser les mécanismes sous-jacents aux accidents du travail. Parmi ces modèles figurent celui de Surry [16], le modèle du "fromage suisse" proposé par Reason [13], ainsi que l'arbre des causes développé par l'INRS [9]. Ces approches, adoptées dans une démarche réactive, ont contribué à une meilleure compréhension des causes et des facteurs déclencheurs des accidents, mais restent essentiellement rétrospectives et présentent des limites [10] : elles sont purement qualitatives, ne prennent pas en compte la dimension temporelle et ne possèdent aucune capacité prédictive.

Les entreprises disposent de données largement sous-exploitées, notamment les remontées terrain dont leur fréquence permet de faire émerger des signaux faibles de risque imminent d'accidents mineurs, voire d'accidents graves. En analysant ces données, il est possible de développer des modèles susceptibles de prévoir la survenue d'accidents du travail. Le concept de pyramide des risques [3][7], populaire dans le domaine de la prévention des accidents, a montré que l'analyse des incidents mineurs (situations dangereuses, quasi-accidents) peut contribuer à prévenir des accidents plus graves.

Cet article propose une méthodologie générique pour exploiter ces données afin de prévoir, à horizon court, la survenue d'accidents au sein d'une entreprise, aussi bien de manière globale que par service. Pour y parvenir, les accidents quotidiens sont traités comme des séries temporelles binaires, modélisées à l'aide de méthodes d'apprentissage automatique. Ces méthodes ont déjà démontré leur efficacité dans le domaine de la prévention des accidents, comme en témoignent les travaux de [5] et [14]. Les prévisions quotidiennes sont également agrégées à l'échelle hebdomadaire et évaluées à l'aide de métriques adaptées pour gérer le déséquilibre des classes [12][4]. Une procédure d'évaluation spécifique aux séries temporelles est mise en œuvre [17][2] afin d'examiner l'efficacité opérationnelle de cette méthodologie dans un contexte de prévention.

Cet article est structuré comme suit. La section 2 décrit la méthodologie proposée, en mettant l'accent sur les modèles et l'évaluation par périodes. Enfin, la section 3 introduit les données réelles utilisées, puis présente et analyse les résultats obtenus.

2 Méthode

2.1 Modélisation des accidents comme une série temporelle binaire

Dans cette étude, nous considérons une série binaire $\{y_t\}_{t=1}^{T}$ décrivant l'évolution quotidienne des accidents. Pour chaque jour $t \in \{1, \ldots, T\}$, la variable binaire $y_t \in \{0, 1\}$ prend la valeur 1 lorsqu'au moins un accident survient à la date t, et 0 dans le cas contraire. Les informations issues des remontées terrain sont incorporées sous forme de covariables C_t.

Pour chercher à estimer la probabilité d'occurrence p_{t+h} d'un accident sur un horizon h, nous distinguons deux grandes catégories de modèles en fonction de la dimension de leurs sorties. Les modèles à sorties scalaires [6] peuvent être entraînés selon une approche DirRec [15], qui combine les caractéristiques des stratégies directe et récursive. En partant d'un historique de d jours, $X_t = (y_{t-1}, \ldots, y_{t-d}, C_{t-1}, \ldots, C_{t-d})$, chaque modèle $f_h(\cdot, \theta)$ est spécifique à un horizon h, mais incorpore également les prédictions des étapes précédentes dans ses entrées. Ainsi, pour chaque horizon $h \in \{1, \ldots, H\}$, nous estimons p_{t+h} par :

$$\hat{p}_{t+h} = f_h(X_t^h, \theta) \in [0, 1],$$

où $X_t^h = (X_t, \hat{p}_{t+1}, \ldots, \hat{p}_{t+h-1})$ inclut l'historique initial X_t ainsi que les prédictions des horizons précédents, et où θ représente les paramètres du modèle à estimer.

Les modèles à sorties vectorielles, tels que ceux reposant sur des architectures de *deep learning* [8][11], génèrent quant à eux des probabilités pour plusieurs horizons simultanément. Dans ce cas, le modèle $f(\cdot, \theta)$ produit un vecteur de probabilités :

$$\hat{\mathbf{p}}_t = \left(\hat{p}_{t+1}, \hat{p}_{t+2}, \ldots, \hat{p}_{t+H}\right) = f(X_t, \theta).$$

Enfin, pour transformer chaque probabilité prédite $\hat{p}_{t+h}$ en une classe binaire, un seuil $\tau \in [0, 1]$ est appliqué. La prédiction binaire $\hat{y}_{t+h}$ est alors donnée par :

$$\hat{y}_{t+h} = 1_{\{\hat{p}_{t+h} \geq \tau\}}$$

2.2 Niveau de risque par période

Dans de nombreuses situations, l'objectif n'est pas de déterminer le moment exact où un évènement surviendra, mais plutôt d'évaluer si une période donnée est globalement exposée ou non. Cette démarche s'apparente, par exemple, à la problématique de la demande intermittente en gestion de stocks [18], où une évolution plus globale est privilégiée (par exemple, la disponibilité en stock sur une période) plutôt que le suivi précis du jour où la demande survient. Pour intégrer cette perspective dans le cas des accidents, nous proposons de partitionner la période d'étude en périodes successives de longueur H. Pour la j-ième période, nous définissons l'ensemble d'indices

$$W_j = \{(j-1) \cdot H + 1, \ldots, j \cdot H\},$$

où $j \in \{1, \ldots, P\}$ avec $P = \lfloor \frac{T}{H} \rfloor$.

Nous introduisons ensuite une variable binaire Y_j, égale à 1 si au moins un accident est observé au cours de cette période, et 0 sinon. La prédiction associée est notée $\hat{Y}_j$. La prédiction binaire et la valeur réelle pour la période j sont données, respectivement, par

$$\hat{Y}_j = \max_{t \in W_j} \hat{y}_t \quad \text{et} \quad Y_j = \max_{t \in W_j} y_t.$$

Dans cette étude, nous considérons $H = 7$. Chaque période correspond donc à une semaine, permettant ainsi une analyse à l'échelle hebdomadaire. Nous nous interressons donc au fait qu'une semaine soit à risque plutôt que de déterminer précisément le jour où au moins un accident pourrait survenir.

2.3 Apprentissage, validation croisée pour séries temporelles et évaluation des modèles

La dimension du paramètre θ varie selon les modèles utilisés. De plus, les modèles considérés dans les expériences présentées ci-dessous font intervenir des hyperparamètres qui leurs sont propres. Pour l'estimation des hyperparamètres des modèles et des paramètres θ associés, nous avons recours à une validation croisée pour séries temporelles (VCST) avec origine glissante [17] [2]. Cette méthode reproduit les conditions d'utilisation réelles des modèles et permet d'évaluer son comportement sur des données hors échantillon. Afin de mesurer les performances des modèles dans un contexte de classes déséquilibrées, nous nous appuyons également sur des métriques telles que le rappel (RA), la précision (PR), le F1-score (F1) et le BA (balanced accuracy en anglais [4]). Ces indicateurs nous fournissent une vision complète de la capacité prédictive de nos modèles testés.

Par construction, comme détaillé dans la section précédente, les données sont agrégées par semaine lors de l'évaluation. Par conséquent, tous les critères pour l'apprentissage et l'évaluation des modèles considérés sont calculés sur les périodes W_j, donc à l'échelle hebdomadaire :

$$\mathrm{RA} = \frac{\sum_{j=1}^{P} 1(Y_j = 1 \wedge \hat{Y}_j = 1)}{\sum_{j=1}^{P} 1(Y_j = 1)}, \quad \mathrm{PR} = \frac{\sum_{j=1}^{P} 1(Y_j = 1 \wedge \hat{Y}_j = 1)}{\sum_{j=1}^{P} 1(\hat{Y}_j = 1)},$$

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{PR} \cdot \mathrm{RA}}{\mathrm{PR} + \mathrm{RA}}, \quad \mathrm{BA} = \frac{\mathrm{RA} + \mathrm{SP}}{2}, \quad \mathrm{SP} = \frac{\sum_{j=1}^{P} 1(Y_j = 0 \wedge \hat{Y}_j = 0)}{\sum_{j=1}^{P} 1(Y_j = 0)} \quad \text{(Spécificité)} .$$

3 Données, résultats et discussion

Cette étude s'appuie sur les données issues d'une entreprise spécialisée dans la gestion des déchets industriels collectées entre janvier 2019 et novembre 2023 dans 21 services. Les accidents dans ces données sont classés en deux catégories : employés internes ou temporaires (EIT) et prestataires ou intervenants externes (PIE). Concernant les EIT, l'analyse se concentre également sur le service principal, dénommé EIT-S1. Les covariables, telles que le nombre de situations dangereuses, proviennent des observations enregistrées lors des visites de sécurité. La période 2019-2022 a été utilisée comme base d'entraînement, tandis que les données postérieures ont servi pour les tests. Le tableau 1 résume les principales caractéristiques des données.

		Entraînement		Test	
Séries	T	**Distribution**	**Nombre de SD**	**Distribution**	**nombre de SD**
EIT	1401	0 :863 / 1 :231	1066	0 :262 / 1 :46	329
EIT-S1	1331	0 :1041 / 1 :50	354	0 :293 / 1 :14	117
PIE	1376	0 :970 / 1 :103	86	0 :269 / 1 :35	216

TABLEAU 1 – Distribution des données par séries, et nombre de situations dangereuses (SD) pour les données d'entraînement et de test.

Le tableau 2 présente les performances des différents modèles sur les trois séries analysées. Le modèle « Naïf » se base sur une approche simple, consistant à répéter les valeurs des mêmes jours de la semaine précédente pour prédire les jours à venir. En termes de balanced accuracy (BA), le modèle LSTM surpasse les autres modèles sur les trois séries. Il se distingue par sa capacité à détecter efficacement les périodes à risque et sans risque, comme illustré dans la Figure 1. En revanche, les modèles de machine learning traditionnels présentent des performances limitées, proches de celles du modèle naïf, avec une BA à peine supérieure à 50% pour la plupart. Ces faibles résultats pourraient être liés à une mauvaise calibration des probabilités, rendant difficile l'identification de seuils optimaux pour distinguer les classes. De plus l'approche DirRec, bien qu'hybride, amplifie les erreurs de prédiction à chaque horizon. Si les premières étapes contiennent des erreurs ou des imprécisions, celles-ci se propagent aux horizons suivants, ce qui pourrait réduire l'efficacité de ces modèles à prévoir les accidents.

	EIT				EIT-S1				PIE			
Modèle	**RA**	**PR**	**F1**	**BA**	**RA**	**PR**	**F1**	**BA**	**RA**	**PR**	**F1**	**BA**
Naïve	0.64	0.67	0.65	0.37	0.15	0.15	0.15	0.39	0.58	0.58	0.58	0.51
Régression logistique	0.93	0.74	0.83	0.51	0.85	0.31	0.46	0.51	0.71	0.55	0.62	0.46
LDA	0.42	0.76	0.54	0.53	0.31	0.27	0.28	0.46	0.92	0.56	0.70	0.49
Arbre de décision	1	0.74	0.85	0.50	1	0.31	0.47	0.50	0.96	0.56	0.70	0.48
SVC	0.9	0.74	0.81	0.50	1	0.32	0.49	0.53	0.58	0.54	0.56	0.46
Forêt aléatoire	0.93	0.76	0.84	0.56	0.07	0.33	0.12	0.50	1	0.57	0.72	0.50
XGBoost	0.29	0.6	0.39	0.37	0.69	0.26	0.38	0.41	1	0.57	0.73	0.50
LightGBM	0.29	0.64	0.4	0.41	0.33	0.29	0.38	0.48	0.83	0.54	0.65	0.44
HistGradientBoosting	0.45	0.67	0.54	0.41	0.54	0.29	0.38	0.48	0.87	0.58	0.7	0.52
MLP	0.91	0.81	0.86	0.51	0.08	0.08	0.08	0.34	0.33	0.47	0.39	0.42
LSTM	**0.76**	**0.90**	**0.82**	**0.69**	**0.77**	**0.83**	**0.8**	**0.85**	**0.75**	**0.78**	**0.76**	**0.74**

TABLEAU 2 – Évaluation des modèles sur les 3 séries analysés.

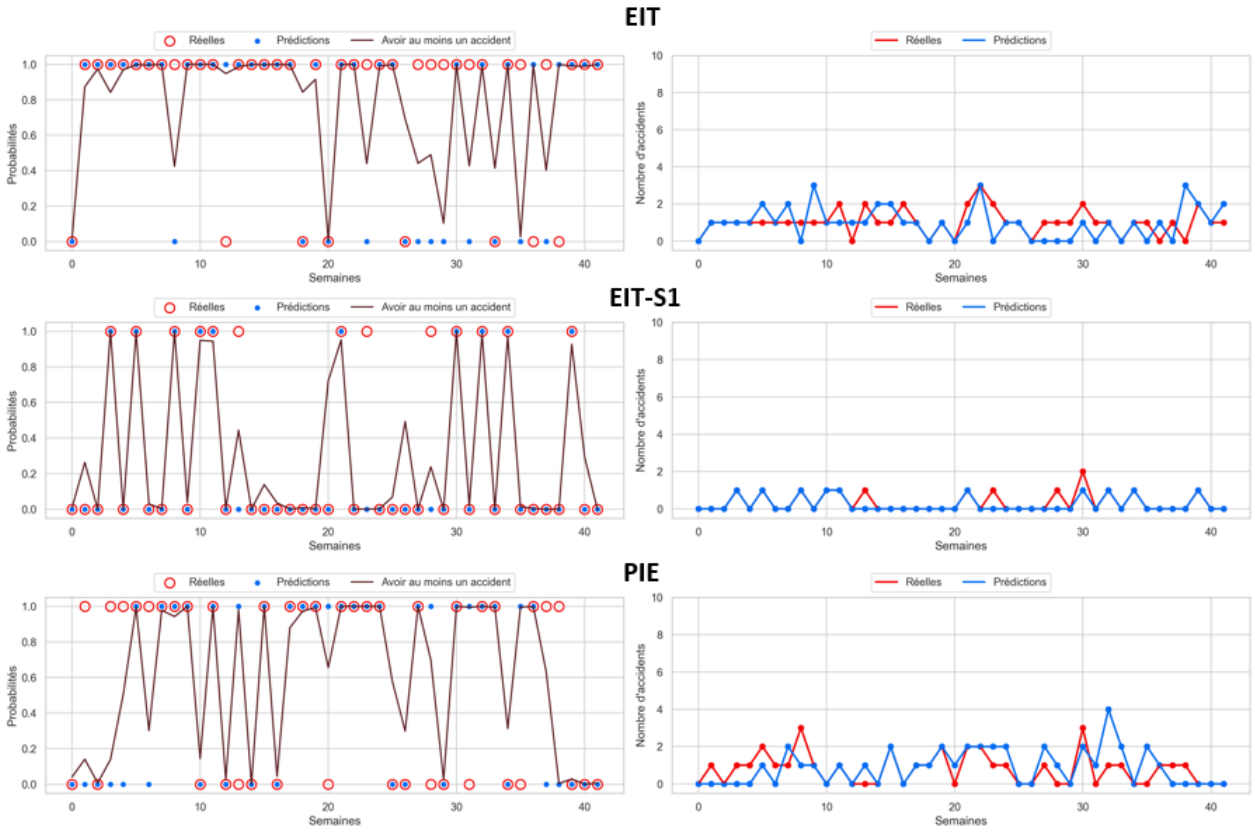


FIGURE 1 – Ajustement des modèles (à gauche) et nombre d'accidents observés et prévus par semaine (à droite) pour les séries EIT, EIT-S1 et PIE.

Bibliographies

[1] Assurance Maladie - Risques professionnels (2024). Rapport annuel 2023 de l'assurance maladie - risques professionnels eléments statistiques et financiers.

[2] Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191 :192–213.

[3] Bird, F. E. (1969). *Management Guide to Loss Control*. Institute Press.

[4] Brodersen, K., Ong, C., Stephan, K., and Buhmann, J. (2010). The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124. IEEE.

[5] Cavalcanti, A. L., Pereira, J. M., and Souza, R. L. (2023). Construction accident prediction using machine learning : A systematic review. *Automation in Construction*, 145 :104621.

[6] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*, volume 2. Springer.

[7] Heinrich, H. W. (1931). *Industrial Accident Prevention : A Scientific Approach*. McGraw-Hill, 1st edition.

[8] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8) :1735–1780.

[9] INRS (1970). Analyse des causes des accidents : Méthodologie et applications. In *Proceedings of the INRS Conference*.

[10] Larouzee, J. and Le Coze, J.-C. (2020). Good and bad reasons : The swiss cheese model and its critics. *Safety Science*, 126 :104660.

[11] Lim, B. and Zohren, S. (2021). Time-series forecasting with deep learning : A survey. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 379(2194) :20200209.

[12] Luque, A., Carrasco, A., Martín, A., and De Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91 :216–231.

[13] Reason, J. (1990). *Human Error*. Cambridge University Press.

[14] Sarkar, S., Reddy, K. S., and Noh, S. (2020). Machine learning models for workplace accident risk prediction. *Safety Science*, 127 :104703.

[15] Sorjamaa, A. and Lendasse, A. (2006). Time series prediction using dirrec strategy. In Verleysen, M., editor, *European Symposium on Artificial Neural Networks, ESANN 2006*, pages 143–148, Bruges, Belgium.

[16] Surry, D. (1969). A model for the analysis of accident sequences. *Accident Analysis & Prevention*, 1(1) :19–23.

[17] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy : An analysis and review. *International Journal of Forecasting*, 16(4) :437–450.

[18] Wallström, P. and Segerstedt, A. (2010). Evaluation of forecasting error measurements and techniques for intermittent demand. *International Journal of Production Economics*, 128(2) :625–636.

Prix SD-SFdS et Session Enseignement

SARP.moodle : creer des questions en grand nombre pour Moodle avec R. Exemple d'application a l'apprentissage des statistiques

Emmanuel Curis [1,2], Marie Verbanck [1] & Virginie Lasserre [1]

[1] UR 7537 BioSTM, faculté de pharmacie de Paris, université Paris Cité, Paris, France
[2] Service d'hématologie, hôpital Lariboisière, AP-HP, Paris, France
emmanuel.curis@u-paris.fr ; marie.verbanck@u-paris.fr ; virginie.lasserre@u-paris.fr

Résumé

En statistique, comme dans d'autres disciplines, les exercices d'application sont indispensables pour acquérir un savoir-faire. Réaliser plusieurs fois le même exercice d'application facilite la consolidation des compétences, à condition de l'accompagner d'une correction et, éventuellement, de conseils. Toutefois, ces compétences ne doivent pas être confondues avec une mémorisation de la bonne réponse ou une suite de tentatives pour trouver par tâtonnement la bonne combinaison de réponses. De ce fait, l'exercice, tout en restant fondamentalement le même, doit varier dans son énoncé à chaque tentative — et la correction doit suivre. Ainsi, pour la compétence « calculer la moyenne d'une série de valeurs », à chaque tentative les valeurs, mais aussi la taille de l'échantillon, doivent changer. De la même façon, dans un contexte d'examen en ligne, disposer d'exercices d'identique difficulté mais différents, par exemple pour limiter les fraudes, est nécessaire.

Parmi les plateformes d'apprentissage, Moodle est un outil adapté à cette double démarche pédagogique, grâce à sa base de questions organisée en « catégories » et la possibilité de tirer au sort une question dans une catégorie pour réaliser un exercice (activité « test » de Moodle), si l'on associe une compétence identifiée à une catégorie. Cependant, la création de questions, en particulier complexes, est rapidement fastidieuse surtout lorsqu'il s'agit de générer de très nombreuses variantes d'une même question. Or les deux situations évoquées créent rapidement le besoin de variantes en très grand nombre d'un même exercice, en se fondant sur une trame d'exercice couplée à une génération aléatoire de ses variantes.

Pour pallier cette difficulté, nous avons développé le *package* R SARP.moodle qui permet d'utiliser les possibilités du logiciel R pour générer aléatoirement des séries de valeurs et de s'affranchir du format XML complexe décrivant les questions pour Moodle. Ce *package* permet de générer un nombre quasi-illimité de variantes d'un même exercice avant import dans Moodle.

Cette présentation illustrera plusieurs utilisations de ce *package* pour générer de telles collections d'exercices, avec un niveau de complexité pouvant être élevé : questions calculatoires, graphiques (comme légender une figure), de démarche de résolution… Quoique ces exemples soient issus des statistiques, cette démarche de simulation d'exercices se généralise à de nombreuses disciplines, en particulier scientifiques.

Nous présenterons aussi comment cet outil permet à des enseignants ne souhaitant pas apprendre R, mais souhaitant créer de nombreuses questions, de préparer leurs questions dans un tableur en respectant un gabarit, souple et simple, permettant une flexibilité et un gain de temps considérables.

Mots-clés. Questions aléatoires ; Moodle ; SARP.moodle ; R

Abstract

In statistics, as in other subjects, practical exercises are essential for acquiring skills. Carrying out the same exercise several times makes it easier to consolidate skills, as long as it is accompanied by a correction and, if necessary, advice. However, these skills should not be confused with memorising the right answer or trying to find the right combination of answers by trial and error. As a result, the exercise, while remaining fundamentally the same, must vary in its wording with each attempt - and the correction must follow. Thus, for the skill 'calculating the mean of a series of values', the values, and also the size of the sample, must change with each attempt. Similarly, in the context of an online exam, it is necessary to have exercises that are identical in difficulty but different, for example to limit cheating.

Among learning platforms, Moodle is a tool suited to this dual pedagogical approach, thanks to its database of questions organised into 'categories' and the possibility of drawing a question from a category to complete an exercise ('test' activity in Moodle), if an identified skill is associated with a category. However, creating questions, especially complex ones, can quickly become tedious, especially when you have to generate a large number of variants of the same question. The two situations mentioned above quickly create the need for a very large number of variants of the same exercise, based on an exercise framework coupled with random generation of its variants.

To overcome this difficulty, we have developed the R package SARP.moodle, which uses the capabilities of R software to randomly generate series of values and does away with the complex XML format used to describe Moodle questions. This package can be used to generate a virtually unlimited number of variants of the same exercise before importing it into Moodle.

This presentation will illustrate several uses of this package to generate such collections of exercises, with a potentially high level of complexity: arithmetic questions, graphical questions (such as how to caption a figure), questions on how to solve a problem, etc. Although these examples come from statistics, this approach to simulating exercises can be applied to many disciplines, particularly science.

We will also show how this tool can be used by teachers who do not wish to learn R, but who do wish to create a large number of questions, to prepare their questions in a spreadsheet using a simple, flexible template that offers considerable flexibility and time savings.

Keywords. Random questions ; Moodle ; SARP.Moodle ; R

En statistiques, comme dans de nombreuses disciplines, la résolution d'exercices est une étape indispensable de l'apprentissage. Elle permet l'acquisition de compétences pratiques, aussi bien techniques (« savoir calculer la moyenne d'un échantillon », « savoir calculer l'intervalle de confiance d'une probabilité »…) que d'interprétation de résultats. Souvent cependant, faire une seule fois l'exercice associé à la compétence n'est pas suffisant. Ainsi, si le calcul d'une moyenne est erroné, pouvoir recommencer l'exercice jusqu'à ce que le calcul soit correct permet de continuer à s'entraîner et ainsi d'acquérir la compétence souhaitée, avec la possibilité de donner des pistes sur les erreurs éventuelles. Mais une fois la réponse correcte donnée, refaire le même exercice exactement n'a que peu d'intérêt. Pour que trouver la solution ne soit pas un simple apprentissage par cœur de cette réponse, ni une déduction par tâtonnement, il est donc nécessaire de disposer d'une collection d'exercices de même nature, mais tous différents, que l'étudiant pourra faire jusqu'à ce qu'il ait acquis la compétence souhaitée.

Depuis la généralisation des plateformes pédagogiques en ligne, telle Moodle, il n'y a pas de réelle limite technique au nombre de variantes : en effet, le système peut proposer à l'étudiant de refaire l'exercice autant de fois qu'il le souhaite et, à chaque nouvelle tentative, un nouvel énoncé sera proposé. Par ailleurs, la correction peut, dans une certaine mesure, être automatisée. Le travail autour de ces exercices est donc moindre, puisque l'étudiant peut les faire quand il le souhaite, sans intervention de l'enseignant… une fois les questions générées.

Moodle propose quelques types de questions avec génération aléatoire de valeurs et calcul de la correction, qui permettent en ne créant qu'une seule question de base de générer en pratique un grand nombre de variantes du même exercice. Par exemple, avec la trame « Calculez la moyenne de ces 5 valeurs : x_1, x_2, x_3, x_4 et x_5 », les valeurs x_1 à x_5 seront tirées au sort. Cependant, les possibilités de ces questions restent limitées, du moins avec les types de question de base : ainsi, la taille de l'échantillon sera fixe ; il est quasi-impossible de calculer une médiane…

Moodle permet par ailleurs d'importer des questions générées au préalable. Cela ouvre la possibilité d'utiliser un langage de programmation pour automatiser la création des variantes d'une même question, avec une souplesse bien plus grande. Néanmoins cette approche comporte deux difficultés :

— d'une part, il faut connaître un langage de programmation ;
— d'autre part, il faut connaître le format de fichier utilisé par Moodle pour importer ses questions.

Dans le monde statistique, le langage R est très largement utilisé, et l'on peut donc au moins dans un premier temps considérer le premier point acquis. En revanche, le second point est plus délicat. Moodle accepte plusieurs formats, mais pour profiter de toutes les fonctionnalités de Moodle — et en particulier pouvoir insérer des images dans des questions (« Quelle est la boîte à moustaches représentant cet échantillon ? ») ou créer des questions avec plusieurs champs de réponse (« Calculez la médiane, les quartiles, l'étendue de cet échantillon. ») —, seul le format XML peut être utilisé. Ce format est assez complexe et son apprentissage une réelle difficulté.

C'est afin de remédier à cette difficulté que nous avons développé pour R le *package* SARP.moodle, qui permet de créer une base de questions au format XML de Moodle sans avoir à connaître la syntaxe de ce format. Ainsi, l'utilisateur désirant créer des questions pour Moodle a juste à apprendre un petit nombre de fonctions R, documentées dans le *package* — disponible sur le CRAN —, sans se préoccuper de la syntaxe Moodle. Nous allons illustrer dans la suite de cet article les principales possibilités de ce *package* et son fonctionnement ; nous conclurons en montrant comment ce *package* permet même de s'affranchir de la connaissance du langage R, pour créer des questions en grand nombre.

1. Principe général de l'utilisation

Afin d'éviter tout conflit de nom de fonction avec d'autres bibliothèques existantes, toutes les fonctions du package accessibles à l'utilisateur se terminent par le suffixe .moodle.

Chaque question dans Moodle comporte un titre (qui sert essentiellement à l'identifier dans la base de question Moodle, mais ne sera pas visible par l'étudiant), un type, un énoncé, une ou plusieurs réponses et des commentaires qui seront, éventuellement, affichés après que l'étudiant aura validé sa réponse et peuvent donc servir d'éléments de correction. Le type de la question permet à Moodle de savoir quels sont les éléments attendus pour la réponse. Ainsi, le type « QCM » est associé à une liste de réponses possibles parmi lesquelles l'étudiant devra en cocher une ou plusieurs ; à chaque réponse est associée une indication sur le fait qu'elle soit correcte, partiellement correcte ou incorrecte. En revanche, le type « Numérique » est associé à une réponse ouverte, mais qui ne peut être qu'un nombre. Il existe une dizaine de types de questions différents

Chacun de ces types de question peut être créé, dans SARP.moodle, à l'aide d'une fonction dédiée qui accepte comme arguments, *a minima*, le titre et l'énoncé de la question, ainsi qu'une indication éventuelle de temps conseillé à l'étudiant pour répondre à la question et qui sera automatiquement ajoutée à la fin de la question. Les autres informations nécessaires, plus spécifiques du type de la question, correspondent à des arguments supplémentaires de la fonction. Le tableau 1 précise, pour chaque type de question, le nom de la fonction à appeler.

Question	Fonction *	Remarque
Vrai-Faux	vrai_faux	Deux réponses possibles, choix unique
QCM ou QCU	qcm	
QROC	qroc	Sensible à la casse ou non
Numérique	numerique	Une tolérance sur la réponse peut être définie
Glisser-déposer	glisser_deposer	Étiquettes à glisser sur une image (régions identifiées)
	legender_image	Textes à placer sur une image (sans indication visuelle)
	glisser_textes	Morceaux de textes à glisser dans un texte à trous
« *Cloze* »	question_libre	Seul type de question avec plusieurs champs de réponse possibles
Rédactionnelle	question_ouverte	Pas de correction automatique ; permet un dépôt de document
Description	description	« Question » sans réponse attendue, pour donner une consigne par exemple

Tableau 1 — Principales questions pouvant être créées via SARP.moodle

* Le suffixe .moodle est omis, pour plus de lisibilité

Toutes les questions sont créées dans un unique fichier XML, à importer ensuite dans Moodle. Ce fichier est initialisé, avant création de la toute première question, par un appel à la fonction debuter_xml.moodle() qui nécessite comme seul argument obligatoire le nom du fichier à créer. Lorsque toutes les questions sont créées, le fichier XML est clôturé par un appel à la fonction finir_xml.moodle().

Le nombre de questions générées pouvant vite devenir très important, il est important de bien structurer la banque de questions dans Moodle, pour s'y retrouver plus facilement, mais surtout afin de garantir que le tirage au sort d'une question dans la base donne bien une variante de la question souhaitée et non n'importe quelle question de la banque. Pour cela, la banque de question de Moodle est structurée en *catégories*, que l'on peut considérer être le pendant d'une arborescence de dossiers et sous-dossiers, les questions étant alors l'analogue des fichiers. Dans SARP.moodle, la catégorie à laquelle appartiendront les questions peut être définie par un appel à la fonction categorie.moodle(). D'un point de vue pédagogique, il est pratique d'assimiler une catégorie à une compétence.

1.1. Amélioration des énoncés

Les énoncés des questions, dans Moodle, peuvent être mis en forme à l'aide de balises HTML, incluant éventuellement des styles au format CSS. Pour certaines mises en formes fréquentes — par exemple d'un nombre, d'un échantillon, d'une sortie du logiciel R… —, SARP.moodle propose des fonctions qui préparent le code HTML/CSS nécessaire. Le résultat de ces fonctions est toujours une chaîne de caractères qui peut être concaténée à celles formant l'énoncé.

Un cas particulier extrêmement fréquent est celui des images qui peuvent être insérées dans un énoncé (y compris les réponses ou les commentaires). SARP.moodle permet, grâce à la fonction lier_image.moodle(), à la fois de créer le code HTML dans l'énoncé pour que l'image s'affiche et de convertir l'image de sorte qu'elle soit présente dans le fichier XML — ainsi, il suffit d'import un unique fichier dans Moodle, facilitant le travail de création et de gestion de la banque de questions.

2. Exemples d'utilisation

Nous allons maintenant illustrer les possibilités du *package* au travers de trois exemples, de difficulté croissante. Pour chacun, le but souhaité sera présenté et les points de réflexion pédagogiques ou techniques évoqués, puis le code R nécessaire et enfin une capture d'écran d'une question ainsi créée, telle qu'elle apparaîtrait à un étudiant souhaitant s'entraîner ou devant passer l'examen. À chaque fois, on notera N le nombre de variantes souhaitées et dans tous nos exemples, seules 10 variantes seront créées.

1.1. Calcul de la moyenne d'un échantillon

Incontournable des statistiques descriptives, savoir calculer la moyenne d'un échantillon de valeurs est typiquement un exercice qu'un étudiant devrait pouvoir faire tant qu'il n'y arrive pas, avec un échantillon différent à chaque fois. Le but est donc de créer un exercice dans Moodle qui ressemble à la capture d'écran de la figure 1, les éléments devant varier à chaque tentative étant encadrés en rouge.

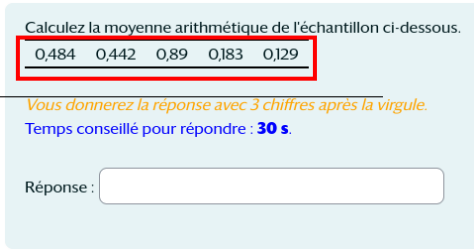


Figure 1 – Exemple de question demandant de calculer une moyenne. Les éléments en rouge doivent varier à chaque tentative.

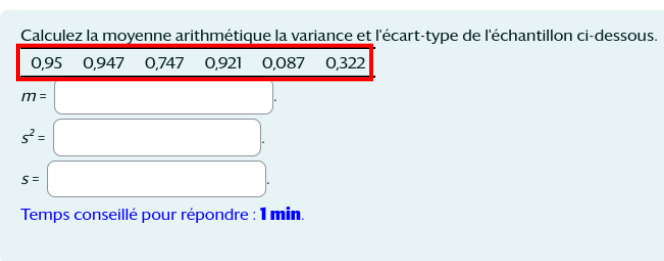
Générer les variantes est, conceptuellement, simple : il suffit de tirer au sort les valeurs, dans une loi quelconque, et de précalculer le résultat pour comparaison avec la valeur fournie par l'étudiant. Plusieurs précautions doivent cependant être prises pour que les questions soient utilisables en pratique. Du point de vue de la question elle-même, les tailles d'échantillon doivent varier tout en ayant le même ordre de grandeur (pour que la difficulté reste comparable d'une tentative à l'autre) et il est important de bien prendre en compte les arrondis. Du côté de l'utilisateur, l'affichage de l'échantillon doit permettre de lire toutes les valeurs même sur desécrans de petite taille — la fonction afficher_echantillon.moodle() est dédiée à l'affichage d'un échantillon : elle va convertir un vecteur de valeurs en un tableau présentant ces valeurs ; si le tableau est trop large pour tenir sur l'écran, un défileur horizontal apparaît, grâce à la propriété CSS adéquate. Il est enfin souhaitable d'indiquer à l'étudiant la précision attendue des résultats et, en particulier pour un examen, le temps indicatif nécessaire pour répondre.

Le code R & SARP.moodle permettant de générer les variantes de la question est indiqué ci-après. Les commandes SARP.moodle sont indiquées en bleu ; la partie du code gérant la création des variantes est en rouge. Les commentaires, en vert, explicitent certains aspects.

```
N <- 10                                  # Nombre de variantes d'une question
debuter_xml.moodle( "moyenne.xml" )      # Ouverture du fichier XML, initialisation
categorie.moodle( "Exemples/Moyenne" )   # Questions à placer en catégorie « Exemples », sous-catégorie « Moyenne »
for ( i in 1:N ) {                       # Création des N variantes de la question
    n <- sample( 4:6, 1 )                # Choix de la taille d'échantillon
    x <- round( runif( n ), 3 )          # On ne garde que trois décimales
    m <- mean( x )                       # Moyenne de cet échantillon arrondi
    # Construction de l'énoncé ; on garde trois décimales dans les valeurs affichées
    énoncé <- paste0( "Calculez la moyenne arithmétique de l'échantillon ci-dessous.",
                      afficher_echantillon.moodle( x, n.chiffres = 3 ) )
    # Construction de la question, réponse attendue avec quatre décimales, temps conseillé de 30 s
    numerique.moodle( énoncé, m, n.decimales = 4, temps = 0.5 )
}
finir_xml.moodle()                       # Clôture du fichier XML
```

1.2. Calcul des statistiques descriptives d'un échantillon

L'étape suivante, en apprentissage des statistiques descriptives, est bien sûr le calcul de plusieurs statistiques descriptives, par exemple la moyenne, la variance et l'écart-type, sur un même échantillon — conduisant à un exercice tel que celui illustré en figure 2, à gauche. Pour cela, les questions numériques de Moodle ne suffisent plus, en effet on ne peut demander qu'une seule valeur par question. Par ailleurs, il n'y a pas moyen dans Moodle de lier des questions indépendantes de sorte qu'elles soient tirées au sort simultanément. Il faut donc se tourner vers un autre type de question : les questions dites « *cloze* ». Ces questions permettent d'avoir autant de champs de réponses dans une même question qu'on le souhaite. La syntaxe de ces questions est cependant complexe.



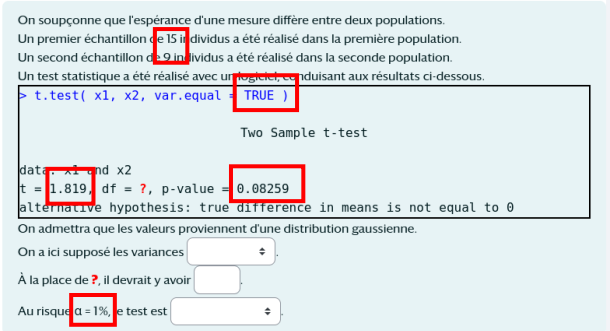


Figure 2 – exemples de questions *clozes*
Les éléments variant à chaque tentative sont encadrés en rouge.

Dans SARP.moodle, pour éviter à l'utilisateur d'apprendre cette syntaxe, nous avons opté pour l'utilisation des listes. Une question *cloze* peut être vue comme une succession de sous-questions, chaque sous-question étant l'association d'un énoncé — le texte avant le champ de réponse — et d'une description du champ de réponse — type de réponse attendu, réponse(s) correcte(s) ou possible(s), contribution à la note de la question totale... La fonction pour créer des questions *cloze* attend donc un vecteur contenant les textes des énoncés et une liste décrivant chaque réponse attendue. Les questions *cloze* permettent une très grande variété de questions et le couplage avec les styles CSS et les balises HTML permet d'organiser les champs de réponse sous des formes très variées, comme le montre la figure 2 (code disponible sur demande).

Pour le premier exemple de la figure 2 le code serait le suivant (en omettant les balises HTML de mise en forme, pour rendre le code plus lisible ; la balise <br /> permet le retour à la ligne) :

```
N <- 10                                 # Nombre de variantes d'une question
debuter_xml.moodle( "moyenne.xml" )     # Ouverture du fichier XML, initialisation
categorie.moodle( "Exemples/Stats des" ) # Questions à placer en catégorie « Exemples », sous-catégorie « Stat des »
for ( i in 1:N ) {
    n <- sample( 4:6, 1 )
    x <- round( runif( n ), 3 )
    m <- mean( x )
    s2 <- var( x ) ; s <- sqrt( s2 )            # Variance et écart-type
    # Construction de l'énoncé ; on garde trois décimales dans les valeurs affichées
    énoncé <- paste0( "Calculez la moyenne arithmétique, la variance et l'écart-type de l'échantillon ci-dessous.",
                      afficher_echantillon.moodle( x, n.chiffres = 3 ) )
    # Construction de la question, réponse attendue avec quatre décimales
    question_libre.moodle( énoncé,              # Début de l'énoncé, « introductif »
                           c(   "m = ",         # Textes avant chaque champ
                                ".<br />s2 = ",
                                ".<br />s = " ),
                           ".",                 # Texte après le dernier champ de réponse
                           list( m, s2, s ),    # Les réponses attendues
                           # Réponse attendue sous forme d'une valeur numérique, pour les trois cas
                           c( "NUMERICAL", "NUMERICAL", "NUMERICAL" ), temps = 1 )
}
finir_xml.moodle()                      # Clôture du fichier XML
```

1.3. Légendage d'une image

Nous présentons ici une question montrant plusieurs densités de probabilité de gaussiennes ; l'étudiant doit attribuer à chacune de ces courbes la gaussienne associée, en faisant glisser une étiquette sur la courbe (figure 3). Le graphe des densités est généré directement en R et le *package* se charge de le traiter pour la conversion en question Moodle au format XML.

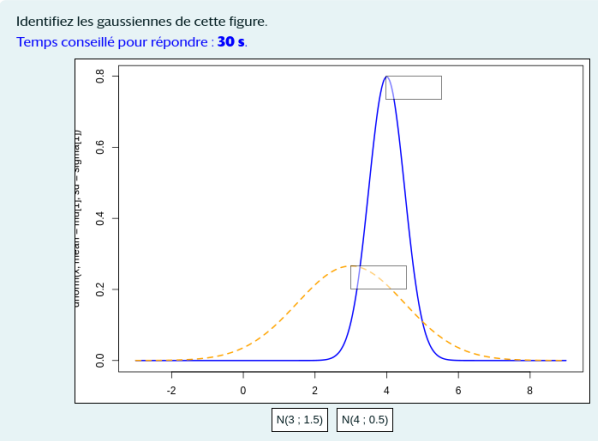


Figure 3 – Exemple de question demandant d'identifier deux gaussiennes.
Les deux gaussiennes changent à chaque tentative.

```
# Fonction construisant la figure des gaussiennes (deux ici)
figure <- function( ) {
    mu <- sample( 1:4, 2 )                  # Espérances
    sigma <- sample( c( 0.5, 1, 1.5 ), 2 )  # Écarts-types
    x <- seq( from = min( mu - 4 * sigma ),
              to = max( mu + 4 * sigma ), length.out = 1000 )
    par( mar = c( 2, 2, 0, 0 ) + 0.5 )      # Pas de marge
    légendes <- paste0( "N(", lapply( mu, afficher_nombre.moodle ), " ; ", sigma, ")" ) # Les étiquettes à glisser/déposer
    # Tracé des deux densités de probabilité, en bleu continu et orange pointillé
    plot(   x = x, y = dnorm( x, mean = mu[ 1 ], sd = sigma[ 1 ] ),
            ylim = c( 0, 1 / ( sqrt( 2 * pi ) * min( sigma ) ) ), type = "l", lwd = 2, col = "blue", ylab = "" )
    lines( x = x, y = dnorm( x, mean = mu[ 2 ], sd = sigma[ 2 ] ), lwd = 2, col = "orange", lty = 2 )
    # Les zones de réponse (où doivent être glissées les étiquettes)
    zones <- data.frame( X = mu, Y = 1 / ( sqrt( 2 * pi ) * sigma ), # Position de ces zones (haut du pic)
```

```
                    Textes = légendes )          # Les textes des réponses
    zones                                        # La fonction renvoie la liste de ces zones
}
N <- 10                                          # Nombre de variantes d'une question
debuter_xml.moodle( "ddp_Gauss.xml" )            # Ouverture du fichier XML, initialisation
categorie.moodle( "Exemples/ddp gauss" )         # Questions à placer en catégorie « Exemples », sous-catégorie « ddp gauss »
for ( i in 1:N ) {
    glisser_deposer.moodle( "Identifiez les gaussiennes de cette figure.", f.creer_figure = figure, temps = "30s" )
}
finir_xml.moodle()                               # Clôture du fichier XML
```

3. Conversion d'un gabarit de questions

Dans certains contextes, on peut souhaiter créer de nombreuses questions, sans qu'elles soient nécessairement des variantes l'une de l'autre, par exemple pour créer une banque de QCM. L'approche présentée auparavant est dans ce cas peu avantageuse, puisque, les questions n'étant pas des variantes, chaque question doit être traitée séparément. Entrer les questions une par une dans Moodle n'est cependant pas très pratique non plus, à partir d'une dizaine de questions.

C'est la raison pour laquelle SARP.moodle propose une autre approche pour cette situation : une fonction qui convertit une banque de questions préparée dans un tableur, suivant un gabarit prédéfini pensé pour être le plus simple possible. Cette conversion peut se faire à partir d'un fichier au format CSV (csv.moodle), Excel (xlsx.moodle) ou Libre Office (ods.moodle).

Brièvement, ce gabarit comporte deux colonnes obligatoires : l'énoncé de la question et la ou les réponses à cet énoncé. Chaque réponse doit être sur une ligne ; les questions sont différenciées (en l'absence de questions *cloze*) par les lignes d'énoncé complétées : tant que la ligne est vide dans la colonne « énoncé », la réponse proposée est une réponse alternative à la même question (figure 4).

titre	Question	Réponse	note	type	temps
Pour tirage au sort dans la compétence	L2/01-moyenne				
Q01 - Moyenne (1)	Une expérience a conduit à l'échantillon ci-dessous 17,25 16,65 15,05 16,50 Quelle est la moyenne arithmétique de cet échantillon ?	16,700			1,5
Q02 - Moyenne (2)	Une expérience a conduit à l'échantillon ci-dessous 11,6 13,4 11,1 10,0 12,9 Quelle est la moyenne arithmétique de cet échantillon ? 	11,800			1,5
Q03 - Moyenne (3)	Une expérience a conduit à l'échantillon ci-dessous 16 18 15 20 19 14 Quelle est la moyenne arithmétique de cet échantillon ?	17,000			1,5
Pour tirage au sort dans la compétence	L2/02-médiane				
Q04 - Médiane (1)	Une expérience a conduit à l'échantillon ci-dessous 17,25 16,65 15,05 16,5 Quelle est la médiane de cet échantillon ?	16,65	1	QCU	1
		16,50	0		
		17,00	0		
		16,00	0		

Figure 4 – exemple de fichier de questions structuré dans un tableur selon le gabarit SARP.moodle
Cet exemple permet de construire deux catégories, la première contenant trois questions à réponse numérique, la seconde une question à choix unique parmi quatre propositions.

Ce gabarit peut être enrichi de colonnes additionnelles pour plus de souplesse : titre de la question, temps conseillé, indication sur le type de question souhaité (pour distinguer un QCM d'un QCU, par exemple), note de la question… Il est aussi prévu pour permettre l'insertion d'images, grâce aux balises @@ encadrant le nom du fichier contenant l'image.

4. Conclusion

Notre *package* SARP.moodle, librement disponible sur le CRAN, permet de créer facilement de nombreuses variantes d'une même question en R, sans apprentissage de la syntaxe de Moodle. Il permet aussi la conversion en masse de questions entrées dans un fichier de tableur, format qui permet un travail collaboratif pour la création des questions même lorsqu'une partie des enseignants ne connaît pas R.

Il s'enrichit progressivement de nouvelles fonctionnalités, au fur et à mesure des échanges avec les utilisateurs. Ainsi, l'un des derniers développements, en cours de finalisation, réalisé à la demande d'enseignants en anatomie, est la possibilité de créer des images à annoter, les annotations se

Figure 5 – Question demandant d'annoter une image.

Prix SD/SFdS : Mise en place d'une base de données pour la gestion des approvisionnements en matières premières de LOEUF

Anaëlle Durand*[1]

[1]Université Bretagne Sud – Master Science des données – France

Résumé

LOEUF est une entreprise grandissante qui a un besoin de fiabiliser ses données et ses prévisions
en matière d'approvisionnement d'œufs.

L'objectif de l'alternance est de répondre à ce besoin en modélisant la ponte avec un modèle additif

généralisé afin d'avoir des estimations précises. Ce modèle donnera lieu à une table dans une base de

données qui permettra, à partir du parc actuel de LOEUF, de prédire le nombre d'œufs entrant dans l'usine

pour chaque semaine.

Ce projet nécessite plusieurs acteurs qui interviennent à différents moments du projet. Nous avons

d'abord la partie collecte des données dans Arcadia géré par le pôle Amont. La partie " flux entre Cognos

et Arcadia " géré par LOEUF et le siège LDC et la dernière partie " construction de tableaux de bords géré

par LOEUF.

Les prévisions actuelles sont suffisantes pour le moment, mais elles prennent du temps à être réalisé

et sont de moins en moins fiable avec l'arrivée de nouveaux élevages. La création de l'outil devient
indispensable avec la croissance de l'entreprise.

*Intervenant

LE TUTORAT PAR LES PAIRS DANS L'ENSEIGNEMENT SUPÉRIEUR À ACCÈS LIBRE : PREUVES DE SON EFFICACITÉ DANS LES UNIVERSITÉS BELGES FRANCOPHONES

Antoine Platteau[1], Philippe Emplit[2], Dorothée Baillet[3] & Catherine Dehon[4]

[1] *TIMES[2], Université libre de Bruxelles (BE), antoine.platteau@ulb.be*
[2] *OPERA-Photonics, Université libre de Bruxelles (BE), philippe.emplit@ulb.be*
[3] *CRSE, Université libre de Bruxelles (BE), dorothée.baillet@ulb.be*
[4] *ECARES, Université libre de Bruxelles (BE), catherine.dehon@ulb.be*

Résumé. Face aux taux d'échec élevés dans l'enseignement supérieur, de nombreux programmes de soutien à l'apprentissage ont été introduits pour améliorer la réussite des étudiants de première année de bachelier. Cette étude évalue l'impact causal du tutorat par les pairs dans le contexte des universités belges francophones, où l'accès est libre et les frais de scolarité faibles. En utilisant la méthode de *propensity score matching* nous estimons les effets du tutorat sur les performances des étudiants de première année dans trois cours de la Faculté des Sciences Psychologiques et de l'Éducation de l'Université libre de Bruxelles. Pour traiter le biais de sélection, nous appliquons différentes spécifications de matching, incluant le K-nearest neighbour matching, avec et sans remplacement, et en ajustant les ratios (1:1, 1:2, 1:3). Nous utilisons également le caliper matching (0,01, 0,05) pour tester la robustesse des résultats. Ces approches permettent de créer des groupes de traitement et de contrôle comparables sur la base de caractéristiques observables. Les résultats montrent une amélioration significative des performances des étudiants tutorés, en particulier dans le cours de Neuropsychologie, où le tutorat a permis d'augmenter les notes de 2,42 en moyen, sur 20. Cette étude contribue à la littérature en confirmant l'efficacité du tutorat par les pairs dans un contexte d'accès libre et de faibles frais d'inscription, moins étudié dans la littérature internationale. Malgré son efficacité, les taux de participation restent faibles, soulignant la nécessité de développer des stratégies pour accroître l'engagement des étudiants. Les recherches futures devraient se concentrer sur l'optimisation de ces interventions et la promotion de la participation pour maximiser les bénéfices du programme.

Mots-clés. Tutorat par les pairs, Impact causal, Réussite étudiante, Propensity score matching, K-nearest neighbour

Abstract. Faced with high failure rates in higher education, many student learning support programs have been introduced to improve undergraduate student success. This study evaluates the causal impact of peer tutoring in the context of French-speaking Belgian universities, where access is open and tuition fees are low. Using propensity score matching, we estimate the effects of tutoring on first-year student performance in three courses from the Faculty of Psychological and Educational Sciences at the Université libre de Bruxelles. To address selection bias, we apply various matching specifications, including K-nearest neighbour matching (KNN) with and without replacement, and adjusting ratios (1:1, 1:2, 1:3). We also use caliper matching (0.01, 0.05) to test the robustness of the results. These approaches create comparable treatment and control groups based on observable characteristics. The results show a significant improvement in the performance of tutored students, particularly in the Neuropsychology

course, where tutoring led to an average increase of 2.42 points out of 20. This study contributes to the literature by confirming the effectiveness of peer tutoring in an open-access, low-tuition context, which has been less studied in the international literature. Despite its effectiveness, participation rates remain low, highlighting the need to develop strategies to increase student engagement. Future research should focus on optimizing these interventions and promoting participation to maximize the program's benefits.

Keywords. Peer tutoring, Causal impact, Student success, Propensity score matching, K-nearest neighbour

1 Introduction

Garantir un enseignement supérieur de qualité et la réussite des étudiants est une priorité dans les institutions à travers le monde. A cette fin, nombreuses sont les institutions qui mettent en place des dispositifs d'aide aux apprentissages, tels que le tutorat par les pairs, dont les moyens utilisés et leur efficacité varie en fonction du contexte et des politiques institutionnelles. Le tutorat par les pairs a notamment démontré son efficacité pour améliorer les résultats académiques dans des systèmes caractérisés par une sélection à l'entrée et à frais de scolarité élevés, comme c'est le cas dans les pays anglo-saxon (Dawson et al., 2014; Topping, 1996). Cependant, son impact dans des systèmes plus inclusifs, de libre accès, tels que celui de la Belgique francophone, a été moins étudié.

Le système d'enseignement supérieur belge francophone, régi par la Fédération Wallonie-Bruxelles (FWB), repose sur un accès libre et des frais de scolarité faibles (835€/an), favorisant l'inclusivité, mais générant d'autres défis. L'accès à l'université, dans la plupart des facultés, se fait sur base de la simple possession du Certificat d'Enseignement Secondaire Supérieur (CESS). Combiné à l'absence d'examen national commun en fin de secondaire, cela crée des disparités de niveau à l'entrée à l'université. La transition vers l'université est difficile pour beaucoup d'étudiants, ce qui se traduit par un faible taux de diplomation de seulement 21% et un taux d'abandon de 30% en bachelier (Lambert, 2023; OECD, 2021).

Pour aider dans cette transition, le programme d'*Aide à la Réussite* impose aux universités de mettre en place des activités de soutien, sans imposition de format, pour les étudiants de première année de bachelier (Décret FWB, 2004). À l'Université libre de Bruxelles (ULB), les formats couramment mis en place sont des rencontres individuelles et des séances de tutorat par les pairs. Cet article analyse l'impact causal de ce dernier type de soutien dans le cadre du tutorat mis en place en première année du bachelier en sciences psychologiques et de l'éducation. Les séances de tutorat, attachées à un cours, principalement ceux à haut taux d'échec, sont animées par des étudiants de troisième année de bachelier ou de master, ayant obtenu une note d'au moins 16/20 et ayant suivi une formation préalable (Baillet, 2015; Salmon et al., 2009).

Cette étude évalue l'impact causal du tutorat par les pairs, dans le cadre du programme d'*Aide à la Réussite*, sur les performances des étudiants à l'ULB. En analysant trois cours de première année de bachelier en sciences psychologiques et de l'éducation, nous utilisons la méthode du *propensity score matching* pour estimer les effets du programme et prendre en compte le biais de sélection. Cette approche nous permet de démontrer l'efficacité du tutorat dans un contexte d'accès libre et de faibles frais d'inscription, contribuant ainsi au débat international sur l'efficacité des programmes d'aide à l'apprentissage, dans le cas des universités belges francophones.

2 Revue de littérature

Le tutorat par les pairs est un dispositif largement étudié dont les bénéfices ont déjà été démontrés. Des programmes comme le *Supplemental Instruction* (*SI*) aux États-Unis et les *Peer-Assisted Study Sessions* (*PASS*) en Australie ont montré que la participation à ce genre d'activité de soutien à l'apprentissage est associée à de meilleurs résultats académiques. Au travers d'analyses descriptives, de nombreux chercheurs montrent que les participants obtiennent des notes moyennes plus élevées que les non-participants (Dancer et al., 2007; Dawson et al., 2014; Lei et al., 2018). Toutefois, cette corrélation ne prouve pas la causalité, en raison du biais de sélection (Bruffaerts et al., 2011; Endrizzi, 2010).

Pour évaluer l'impact causal du programme sur la réussite, la méthode du *propensity score*

matching est utilisée et compare les étudiants ayant suivi le tutorat avec ceux n'ayant pas participé, en contrôlant les caractéristiques observables (Imbens & Wooldridge, 2009; Rosenbaum & Rubin, 1983). Avec cette méthode, Bowman, Preschel, and Martinez (2023) ont constaté que la participation au SI aux États-Unis améliore les notes de 0,10 à 0,19 points sur une échelle de quatre points, tout en réduisant les taux d'échec de 4 à 6 points de pourcentage. De la même manière, Guarcello et al. (2017) ont observé qu'assister à au moins une séance de *SI* augmentaient considérablement les chances de réussir un cours. En Europe, des études en Espagne, en Allemagne et aux Pays-Bas ont également montré des résultats positifs, bien que leur mise en place et leur évaluation reste moins généralisée (Arco-Tirado et al., 2020; Hardt et al., 2023). Cependant, la plupart des études européennes utilisent des analyses descriptives, ce qui ne permet pas de conclure sur la causalité. D'autres encore utiliser des *randomized controlled trials* (RCT), qui, dans le contexte éducationnel, peuvent poser des préoccupations éthiques, en raison de l'exclusion des étudiants des interventions de tutorat (Dekker et al., 2023).

Malgré les preuves soutenant l'efficacité du tutorat par les pairs, son impact dans des systèmes à libre accès et faibles frais d'inscription reste inconnue. La participation volontaire au programme tutorat par les pairs soulève des préoccupations de biais de sélection, rendant difficile l'évaluation de l'impact du programme. Les universités belges francophones, confrontées à une préparation académique hétérogène des étudiants, font face à des taux d'échec et de décrochage élevés. Cette étude contribue à la littérature internationale en évaluant l'impact causal du tutorat dans un système de libre accès à faibles frais d'inscription, offrant ainsi une compréhension plus nuancée de l'efficacité des programmes de soutien dans l'enseignement supérieur.

3 Données et méthodologie

Cette étude analyse les données d'étudiants de première année inscrits au bachelier en en sciences psychologiques et de l'éducation de la Faculté du même nom de l'ULB, sur quatre années académiques (2013-2016). L'ensemble de données inclut 879 étudiants de première génération. Par étudiant de première génération, on entend un étudiant inscrit pour la première fois dans un programme d'enseignement supérieur, n'ayant jamais été inscrit auparavant dans un autre programme de même niveau. Les étudiants de cet ensemble de données ont suivi tous les cours du bachelier mais nous nous intéressons à trois cours en particulier – Neuropsychologie, Cytologie et Statistiques – en raison de leurs taux d'échec élevés, de la disponibilité des registres de présence détaillés pour les séances de tutorat et de leur méthode d'évaluation similaire (QCM).

Dans notre étude, on s'intéresse à la note obtenue par les étudiants à l'examen de chacun des trois cours. Cette note est mesurée sur une échelle de 0 à 20, où une l'obtention d'une note supérieure ou égale à 10 est considérée comme une réussite. Pour le traitement, une variable binaire est utilisée, prenant la valeur 1 si un étudiant a assisté à au moins une séance de tutorat. Nous contrôlons pour une série de variable telles que le genre, l'éligibilité à une bourse (comme indicateur du statut socio-économique), le parcours scolaire secondaire (cours de latin/grec, mathématiques avancées), le redoublement, l'indice socio-économique du lycée de l'étudiant et le type de diplôme du secondaire. Nous contrôlons également pour l'année.

Dans premier temps, on emploie des régressions multivariées pour comprendre les dynamiques du *dataset*, vérifier l'alignement avec la théorie et avoir une idée de l'influence du tutorat sur la réussite. Dans un second temps, pour estimer l'impact causal du tutorat, nous utilisons la méthode du *propensity score matching*. Cette méthode permet de traiter le biais de sélection en associant les étudiants tutorés à des étudiants non tutorés partageant des

caractéristiques observables similaires. L'effet de traitement (représenté par l'*average treatment effect on the treated* – ATT) mesure l'impact du tutorat sur la performance académique. Pour assurer la robustesse de nos résultats, nous utilisons différentes spécifications de matching, y compris le K-nearest neighbour matching (KNN) avec et sans remplacement, avec différents rapports (1:1, 1:2, 1:3), ainsi que le caliper matching (0,01, 0,05) .

Nous vérifions auparavant deux hypothèses clés nécessaires à la validité du *propensity score matching*. D'une part l'hypothèse d'indépendance conditionnelle (CIA), qui exige, en théorie, l'inclusion de toutes les variables influençant la probabilité de traitement dans l'analyse. L'indépendance entre le résultat et le statut de traitement permet d'attribuer la causalité au traitement, garantissant que l'effet observé provient effectivement du traitement et n'est pas influencé par des variables de confusion. D'autre part, afin de garantir que le matching s'est opéré correctement, il faut s'assurer que les variables sont équilibrées entre les deux groupes, étant donné le *propensity score* (Imbens & Rubin, 2015). Le tout en vérifiant qu'il y a suffisamment de recouvrement dans la distribution des *propensity scores* entre les étudiants traités et les étudiants témoins. Le chevauchement est inspecté visuellement à l'aide de graphiques de distribution des scores de propension afin de confirmer que les deux groupes partagent une région de support commune.

4 Résultats

Une première analyse multivariée nous donne une indication de la valeur du coefficient associé à la variable de tutorat, positif dans notre cas. Bien que ce type d'analyse n'établisse pas de lien causal, il constitue un bon point de départ. Les résultats de l'analyse de régression multivariée montrent que le tutorat par les pairs a un effet significatif et positif sur la performance des étudiants. L'impact le plus fort est observé dans le cours de Neuropsychologie, où le tutorat augmente les notes d'environ 2,42 points ($p < 0,01$). Des effets plus petits, mais positifs, sont observés en Cytologie (0,76 points) et en Statistiques (1,29 points). Ces différences peuvent refléter la nature variée de ces cours et de la façon dont les séances de tutorat sont organisées.

La méthode de *propensity score matching* confirme ces résultats, en fournissant des estimations causales de l'impact du programme. En Neuropsychologie, l'effet de traitement varie de 2,19 à 2,57 points, selon les spécificité du modèle utilisées, et reste significatif au niveau de 1%. Pour les cours de Cytologie et les Statistiques, l'effet de traitement varie respectivement de 1,00 à 1,67 et de 1,16 à 1,78 points, lorsque le coefficient est significatif.

Ensuite sur base de l'effet de traitement fourni par le KNN1 logit matching avec remplacement, nous sommes intéressés à la proportion d'étudiants ayant obtenu une note correspond à une réussite, grâce à leur participation aux séances de tutorat. En Neuropsychologie, 31,9 % des étudiants tutorés ont réussi le cours grâce au tutorat, contre 35,3 % en Cytologie et 26,2 % en Statistiques. Cela veut dire qu'en moyenne, un tiers des étudiants qui participent au tutorat réussit le cours grâce à sa participation. Les deux autres groupes sont d'une part les étudiants qui auraient de toute façon réussit et de l'autre, ceux qui auraient de toute façon raté. Ces résultats suggèrent que bien que le tutorat permette d'améliorer la performance dans tous les cours. Le pourcentage plus faible pour le cours de Statistiques peut s'expliquer par le fait que ce cours est souvent considéré comme le plus difficile par les étudiants de psychologie et celui avec le taux d'échec le plus élevé. En raison des difficultés inhérentes au cours, assister aux séances de tutorat ne suffit pas pour réussir.

Malgré des résultats positifs, dans l'ensemble, les taux de participation restent faibles, allant de 12% en Statistiques à 17% en Neuropsychologie, ce qui limite l'impact possible du

programme. Cependant, ces taux sont cohérents avec les recherches antérieures qui constatent un faible taux de participation à des programmes similaires (Cornelius et al., 2016). Surmonter les obstacles à la participation, tels que la stigmatisation et encourager une participation active régulière, est crucial pour améliorer la portée et l'efficacité du programme.

5 Discussion et implication

Les résultats de cette étude confirment que le tutorat par les pairs est une intervention efficace pour améliorer la performance académique dans un contexte d'enseignement supérieur d'accès libre et de faibles frais d'inscription. À l'inverse des systèmes plus sélectifs, un tel programme dans les institutions d'enseignement supérieur de la FWB permet d'aider à combler les lacunes et l'hétérogénéité des niveaux à l'entrée de l'université. On constate que l'effet de traitement varie selon le cours, avec un impact plus important observé en Neuropsychologie, un cours souvent perçu comme exigeant sur le plan conceptuel. Ce résultat souligne l'importance du soutien académique supplémentaire dans les cours nécessitant une forte compréhension conceptuelle et des compétences en pensée critique. L'effet plus modeste pour le cours de Statistiques renforce l'idée d'un cours plus compliqué pour des étudiants en psychologie. Il est également probable que l'ampleur de l'impact puisse varier en fonction de la structure et du contenu des séances de tutorat.

Cependant, les taux de participation relativement faibles soulèvent des questions sur l'attractivité du programme et les facteurs qui empêchent une participation plus large. Les obstacles potentiels peuvent inclure des contraintes de temps et la stigmatisation associée à la demande d'aide académique. Surmonter ces défis est essentiel pour maximiser la portée du programme et garantir que les étudiants qui en ont le plus besoin puissent en bénéficier.

Les résultats suggèrent également que les futures itérations du programme pourraient être adaptées pour mieux répondre aux besoins de populations étudiantes spécifiques. Par exemple, rendre obligatoire la participation aux séances de tutorat ou offrir une sensibilisation ciblée aux étudiants à risque pourrait améliorer les taux de participation et in fine le taux de réussite. De plus, promouvoir une culture qui normalise le comportement de recherche d'aide pourrait réduire la stigmatisation et encourager davantage d'étudiants à tirer parti des ressources disponibles.

Cette étude présente plusieurs limitations. La principale étant que malgré la réduction du biais de sélection par le matching, cela ne tient pas entièrement compte des caractéristiques non observables, telles que la motivation ou les capacités cognitives, qui peuvent influencer à la fois la participation au tutorat et la performance académique. Dès lors, cela peut affecter l'hypothèse d'indépendance conditionnelle. De futures recherches pourraient surmonter cette limitation en intégrant une proxy pour les données manquantes ou bien utiliser la méthode des variables instrumentales pour contrôler sur le biais de sélection lorsqu'il porte sur des variables non observables.

6 Conclusion

Cet article analyse l'efficacité du tutorat par les pairs, en tant que dispositif de soutien à l'apprentissage, sur la performance académique des étudiants de première année à l'ULB. Grâce au *propensity score matching*, nous démontrons l'impact causal de ce dispositif, dans les trois cours analysés, permettant en moyenne à un tiers des participants d'obtenir des notes suffisantes pour valider le cours. Ces résultats démontrent que ce soutien contribue à la réussite des étudiants dans un contexte de libre accès et de faibles frais d'inscription. Cette étude contribue à la littérature internationale en confirmant la validité des résultats démontrés précédemment

dans des systèmes éducatifs à frais élevés et à sélection à l'entrée.

Nos résultats soulignent l'importance du soutien à l'apprentissage pour combler les inégalités de niveaux à l'entrée du supérieur en FWB. Il est crucial de convaincre les étudiants en difficulté de profiter de ces ressources. Une collaboration étroite décideurs politiques et les institutions est nécessaire pour continuer à développer et améliorer de tels programmes et garantir un enseignement supérieur de qualité et accessible à tous.

7 Références

Arco-Tirado, J. L., Fernandez-Martin, F. D., & Hervas-Torres, M. (2020). Evidence-based peer-tutoring program to improve students' performance at the university. *Studies in Higher Education*, *45*(11), 2190–2202.

Baillet, D. (2015). *Les "Guidances"—Dispositif d'aide à la réussite destiné aux étudiants de BA 1*. Université libre de Bruxelles, Faculté des Sciences Psychologiques et de l'Education.

Bowman, N. A., Preschel, S., & Martinez, D. (2023). Does supplemental instruction improve grades and retention? A propensity score analysis approach. *The Journal of Experimental Education*, *91*(2), 205–229. https://doi.org/10.1080/00220973.2021.1891010

Bruffaerts, C., Dehon, C., & Guisset, B. (2011). Can schooling and socio-economic level be a millstone to a student's academic success? In *ECARES Working paper 2011-016*.

Cornelius, V., Wood, L., & Lai, J. (2016). Implementation and evaluation of a formal academic-peer-mentoring programme in higher education. *Active Learning in Higher Education*, *17*(3), 193–205. https://doi.org/10.1177/1469787416654796

Dancer, D., Morrison, K., & Smith, M. (2007). Measuring the impact of a peer assisted learning program on students' academic performance in econometrics. *The Quantitative Analysis of Teaching and Learning in Higher Education: Forum Proceedings*, 19–42. https://espace.library.uq.edu.au/view/UQ:240565/flexiblelearning.pdf#page=21

Dawson, P., van der Meer, J., Skalicky, J., & Cowley, K. (2014). On the effectiveness of supplemental instruction: A systematic review of supplemental instruction and peer-assisted study sessions literature between 2001 and 2010. *Review of Educational Research*, *84*(4), 609–639. https://doi.org/10.3102/0034654314540007

Décret définissant l'enseignement supérieur, favorisant son intégration dans l'espace européen de l'enseignement supérieur et refinançant les universités, Pub. L. No. 28769, Moniteur Belge 45239 (2004). https://www.gallilex.cfwb.be/document/pdf/28769_003.pdf

Dekker, I., Luberti, M., & Stam, J. (2023). Effects of supplemental instruction on grades, mental well-being, and belonging: A field experiment. *Learning and Instruction*, *87*, 101805. https://doi.org/10.1016/j.learninstruc.2023.101805

Endrizzi, L. (2010). Réussir l'entrée dans l'enseignement supérieur. *Dossier d'actualité de La VST*, *59*, 1–23.

Guarcello, M. A., Levine, R. A., Beemer, J., Frazee, J. P., Laumakis, M. A., & Schellenberg, S. A. (2017). Balancing Student Success: Assessing Supplemental Instruction Through Coarsened Exact Matching. *Technology, Knowledge and Learning*, *22*(3), 335–352. https://doi.org/10.1007/s10758-017-9317-0

Hardt, D., Nagler, M., & Rincke, J. (2023). Tutoring in (online) higher education: Experimental evidence. *Economics of Education Review*, *92*, 102350. https://doi.org/10.1016/j.econedurev.2022.102350

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (1st Edition). Cambridge University Press.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86. https://doi.org/10.1257/jel.47.1.5

Lambert, J.-P. (2023). Les politiques de l'éducation et leurs effets sociétaux. *La Thérésienne*, *1*. https://popups.uliege.be/2593-4228/index.php?id=1682

Lei, H., Cui, Y., & Zhou, W. (2018). Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: An International Journal*, *46*(3), 517–528. https://doi.org/10.2224/sbp.7054

OECD. (2021). *Education at a Glance 2021: OECD Indicators*. OECD Publishing. https://doi.org/https://doi.org/10.1787/b35a14e5-en

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Salmon, D., Baillet, D., Boulvain, M., Cobut, B., Coupremanne, M., Duchâteau, D., Lanotte, A.-F., Dubois, P., Goemaere, S., Noël, B., Houart, M., & Slosse, P. (2009). Construction d'un outil d'évaluation de la qualité des actions d'accompagnement pédagogique. Synthèse d'échanges et d'analyse de pratiques professionnelles en Communauté française de Belgique. *Revue internationale de pédagogie de l'enseignement supérieur*, *25*(2). https://doi.org/10.4000/ripes.252

Topping, K. J. (1996). The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *Higher Education*, *32*(3), 321–345. http://dx.doi.org/10.1007/BF00138870

Processus empiriques –Empirical processes, Statistique des processus

Inference of Hawkes Processes with Variable Length Memory

Sacha Quayle[1] & Anna Bonnet[2] & Maxime Sangnier[3]

[1,2,3] *LPSM, Sorbonne Université, France.*
[1] *quayle@lpsm.paris* [2] *bonnet@lpsm.paris* [3] *sangnier@lpsm.paris*

Résumé. Les processus de Hawkes classiques, avec ou sans inhibition, sont étudiés pour modéliser des phénomènes où l'apparition d'un évènement augmente ou diminue la probabilité d'apparition d'évènements futurs sur une certaine période de temps. Dans le cadre des processus de Hawkes à mémoire variable, la probabilité d'apparition d'un évènement dans un sous-processus donné dépend uniquement de ce qui s'est passé depuis le dernier évènement de ce sous-processus. Nous étudions un modèle plus général, englobant à la fois les modèles classiques et à mémoire variable, et nous proposons des méthodes d'estimation des paramètres par maximum de vraisemblance.

Mots-clés. Processus ponctuel, processus de Hawkes, inhibition, auto-excitation, mémoire variable, intensité conditionnelle, estimation par maximum de vraisemblance.

Abstract. Classic Hawkes processes, with or without inhibition, are studied to model phenomena where the occurrence of one event influences the likelihood of subsequent events over a given period. In the case of variable memory Hawkes processes, the probability of an event in a given subprocess depends only on what has occurred since the last event in that subprocess. We study a more general model, which encompasses both the classic and variable memory models, and we propose parameter estimation methods by maximum likelihood estimation.

Keywords. Point process, Hawkes process, inhibition, self-excitation, variable memory, conditional intensity function, maximum likelihood estimation.

1 Introduction

Hawkes processes generalise Poisson processes by incorporating self-excitation, where each event increases the probability of appearance of future events within a certain time window. First introduced to model earthquakes [Hawkes, 1971], these processes have since been applied in social media analysis [Rizoiu et al., 2017], criminology [Olinde & Short, 2020], and neuroscience [Lambert et al., 2018].

While originally univariate and self-exciting [Hawkes & Oakes, 1974], Hawkes processes have been extended to multivariate settings, allowing subprocess interactions, with inference methods such as maximum likelihood estimation being well-documented [Ogata, 1978]. Recent work also explores inhibitory effects, where events reduce the likelihood of future events.

This model remain less studied due to its non-linear nature, but theoretical results have been established [Brémaud & Massoulié, 1996, Costa et al., 2020], and there is research on maximum likelihood estimation for these processes [Bonnet et al., 2023].

In the previous Hawkes processes, the probability of an event occurring at a given time depends on all previous events. However, some studies suggest a memory reset mechanism may better suit neuronal activity, which led to the introduction of a model with variable length memory [Galves & Löcherbach, 2013, Hodara & Löcherbach, 2017], where event dependencies are limited to those following the last occurrence in a given subprocess. To our knowledge, statistical inference for this model has not yet been explored.

In this communication, we study a generalised model which encompasses both the classic model and the variable length model. For the case of exponential kernels, we derive maximum likelihood estimation methods of the parameters and we illustrate this approach through Python-implemented numerical studies on synthetic data.

2 Mathematical setting

Notations. Let $N = (N^1, ..., N^d)$ be a multivariate point process on $\mathbb{R}_+$. Denote $(T_k^j)_{k\in\mathbb{N}^*}$ the arrival times for each process N^j, and $(T_{(l)})_{l\in\mathbb{N}^*}$ the ordered sequence of arrival times of the process N, with respective associated dimensions $(d_l)_{l\in\mathbb{N}^*}$.

Definition 1. *Let $\mu_i > 0$ and $h_{ij} : \mathbb{R}_+ \to \mathbb{R}$ be measurable functions ($1 \leq i, j \leq d$). A multivariate point process $N = (N^1, ..., N^d)$ with respective conditional intensities $\lambda^1, ..., \lambda^d$ is:*

1. *A Hawkes process (HP) if:* $\lambda^i(t) = \left(\mu_i + \sum_{j=1}^d \int_{[0,t[} h_{ij}(t-s)\,\mathrm{d}N^j(s)\right)^+$.

2. *A variable length memory Hawkes process (VM HP) if:*

$$\lambda^i(t) = \left(\mu_i + \sum_{j=1}^d \int_{[T^i_{N^i(t)},t[} h_{ij}(t-s)\,\mathrm{d}N^j(s)\right)^+.$$

3. *A general variable length memory Hawkes process (GVM HP) if:*

$$\lambda^i(t) = \left(\mu_i + \sum_{j=1}^d \left[\int_{[T^i_{N^i(t)},t[} h_{ij}(t-s)\,\mathrm{d}N^j(s) + \int_{[0,T^i_{N^i(t)}[} \widetilde{h}_{ij}(t-s)\,\mathrm{d}N^j(s)\right]\right)^+.$$

Model (GVM HP) is more general than the others, reducing to Model (HP) when $\widetilde{h}_{ij} = h_{ij}$ and to Model (VM HP) when $\widetilde{h}_{ij} = 0$.

Model assumptions. We suppose that $h_{ij}(t) = \alpha_{ij} e^{-\beta_i t}$ and $\widetilde{h}_{ij}(t) = \widetilde{\alpha}_{ij} e^{-\beta_i t}$, with $\alpha_{ij}, \widetilde{\alpha}_{ij} \in \mathbb{R}$ and $\beta_i > 0$, such that the spectral radius of the matrix:

$$H := (|\max(\alpha_{ij}, \widetilde{\alpha}_{ij})|/\beta_i)_{1 \leq i,j \leq d}$$

is strictly smaller than 1. We therefore denote $\theta = (\mu_i, \alpha_{ij}, \beta_i, \widetilde{\alpha}_{ij})_{1 \leq i,j \leq d} =: (\mu, \alpha, \beta, \widetilde{\alpha})$, and we define the underlying intensities (such that $\lambda_\theta^i(t) = \lambda_\theta^{i*}(t)^+$) by:

$$\lambda_\theta^{i*}(t) = \mu_i + \sum_{j=1}^{d} \left[\sum_{T^i_{N^i(t)} \leq T^j_k < t} \alpha_{ij} e^{-\beta_i (t - T^j_k)} + \sum_{T^j_k < T^i_{N^i(t)}} \widetilde{\alpha}_{ij} e^{\widetilde{\beta}_i (t - T^j_k)} \right].$$

3 Estimation

To estimate θ, we use Maximum Likelihood Estimation (MLE), as done in [Bonnet et al., 2023]. Lemma 3.1 and Proposition 3.1 from the latter article still hold under Model (GVM HP), the only difference is the computation of the intensity values $\lambda_\theta^i(T_{(l)})$, which is done using Proposition 1 below. To compute these intensities recusively, we denote, for any $i \in \{1, ..., d\}$:

$$\eta_\theta^i(t) = \sum_{j=1}^{d} \sum_{T^i_{N^i(t)} \leq T^j_k < t} \alpha_{ij} e^{-\beta_i (t - T^j_k)}, \quad \text{and} \quad \widetilde{\eta}_\theta^i(t) = \sum_{j=1}^{d} \sum_{T^j_k < T^i_{N^i(t)}} \widetilde{\alpha}_{ij} e^{-\widetilde{\beta}_i (t - T^j_k)},$$

and we introduce, for computation purposes: $\widetilde{\eta}^i_{\theta,aux}(t) = \sum_{j=1}^{d} \sum_{T^i_{N^i(t)} \leq T^j_k < t} \widetilde{\alpha}_{ij} e^{-\widetilde{\beta}_i (t - T^j_k)}$.

Proposition 1. *For any $1 \leq i \leq d$:*

$$\forall t \in]T_{(l)}, T_{(l+1)}[, \qquad \lambda_\theta^{i*}(t) = \mu_i + \left[\eta_\theta^i(T_{(l)}) + \alpha_{i,d_l}\right] e^{-\beta_i (t - T_{(l)})} + \widetilde{\eta}_\theta^i(T_{(l)}) e^{-\widetilde{\beta}_i (t - T_{(l)})}, \text{ and}$$

$$\lambda_\theta^{i*}(T_{(l+1)}) = \begin{cases} \mu_i + \left[\widetilde{\eta}_\theta^i(T_{(l)}) + \widetilde{\eta}^i_{\theta,aux}(T_{(l)}) + \widetilde{\alpha}_{i,d_l}\right] e^{-\widetilde{\beta}_i (T_{(l+1)} - T_{(l)})} & i = d_{l+1} \\ \lambda_\theta^{i*}(T^-_{(l+1)}) & i \neq d_{l+1} \end{cases}.$$

By denoting $\ell_\theta(t)$ the log-likelihood [Daley & Vere-Jones, 2003] (Proposition 7.2.III p.232), our estimator is thus $\hat{\theta}_T := \underset{\theta \in \Theta}{\operatorname{argmax}}\ \ell_\theta(T)$.

4 Numerical results

We simulate data using the same parameters μ, α, and β as in Figure 5 of [Bonnet et al., 2023], employing the thinning algorithm and Proposition 1. Two scenarios are considered for the

simulations: Scenario (1) where $\widetilde{\alpha} = \alpha$, corresponding to data from Model (HP), and Scenario (2) where $\widetilde{\alpha} = 0$, corresponding to data from Model (VM HP). For each scenario, the procedure is as follows:

1. Simulate 25 independant realisations ensuring total number of events is always 10,000.
2. Estimate parameters for each realisation under the three models: (HP), (VM HP) and (GVM HP).
3. Compute the average of the 25 parameter estimates for each model: $\widehat{\mu}$, $\widehat{\alpha}$, $\widehat{\beta}$ under Models (HP) and (VM), and $\widehat{\mu}$, $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\widetilde{\alpha}}$ under Model (GVM HP).
4. Using the averaged estimates, perform goodness-of-fit tests on 25 independant realisations applying the time-change theorem [Daley & Vere-Jones, 2003] (Theorem 7.4.I p.258).

To evaluate the quality of the parameter estimates, we analyse the relative squared errors for each group of parameters (μ, α, β and $\widetilde{\alpha}$) for each realisation, by considering vector norms. We then display the average error. For the goodness-of-fit tests, we report the average p-value across the 25 realisations. Results for Scenario (1) and Scenario (2) are summarised in Tables 1 and 2, respectively.

Model Estimation	μ	α	β	$\widetilde{\alpha}$	Average p-value
(GVM HP)	0.033	0.088	0.33	0.16	0.44
(VM HP)	0.038	1.34	0.35	/	0.08
(HP)	0.029	0.057	0.35	/	0.43

Table 1: Average relative squared errors (by considering vector norms) and average p-values for Scenario (1).

Model Estimation	μ	α	β	$\widetilde{\alpha}$	Average p-value
(GVM HP)	0.04	0.1	0.78	0.23	0.47
(VM HP)	0.025	0.072	0.71	/	0.46
(HP)	0.044	1.21	0.82	/	0.09

Table 2: Average relative squared errors (by considering vector norms) and average p-values for Scenario (2).

Model (GVM HP) provides accurate parameter estimates and consistently high p-values across both scenarios, demonstrating its flexibility and ability to fit the data well. Model (HP) performs well when data is from (HP) but fails when the data comes from (VM HP), as reflected by large estimation errors and lower p-values. Similarly, Model (VM HP) fits (VM HP) data accurately but performs poorly for (HP) data. This highlights the limitations of each specific model and the robustness of the generalised Model (GVM HP).

5 Perspectives

The stationarity of general variable-memory Hawkes processes, as well as the conditions on the functions (h_{ij}) and $(\widetilde{h}_{ij})$ to ensure their existence, remain open questions. Additionally, the theoretical properties of the estimators, including their existence, consistency, and asymptotic normality, are yet to be rigorously established. Another important direction for future work is to develop statistical tests to determine whether each subprocess follows classical dynamics or dynamics with variable memory, enabling more precise model selection and validation. Finally, we have yet to apply these methodologies to real neuronal activity data.

References

[Bonnet et al., 2023] Bonnet, A., Herrera, M. M., & Sangnier, M. (2023). Inference of multivariate exponential Hawkes processes with inhibition and application to neuronal activity. *Statistics and Computing*, 33(4), 91. http://arxiv.org/abs/2205.04107.

[Brémaud & Massoulié, 1996] Brémaud, P. & Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *The Annals of Probability*, 24(3), 1563–1588.

[Costa et al., 2020] Costa, M., Graham, C., Marsalle, L., & Tran, V. C. (2020). Renewal in Hawkes processes with self-excitation and inhibition. *Advances in Applied Probability*, 52(3), 879–915. http://arxiv.org/abs/1801.04645.

[Daley & Vere-Jones, 2003] Daley, D. J. & Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*, volume 1 of *Probability and Its Applications (New York)*. Springer-Verlag, 2 edition.

[Galves & Löcherbach, 2013] Galves, A. & Löcherbach, E. (2013). Infinite Systems of Interacting Chains with Memory of Variable Length—A Stochastic Model for Biological Neural Nets. *Journal of Statistical Physics*, 151(5), 896–921. https://arxiv.org/abs/1212.5505.

[Hawkes & Oakes, 1974] Hawkes, A. & Oakes, D. (1974). A Cluster Process Representation of a Self-Exciting Process. *Journal of Applied Probability*, 11, 493–503.

[Hawkes, 1971] Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1), 83–90.

[Hodara & Löcherbach, 2017] Hodara, P. & Löcherbach, E. (2017). Hawkes processes with variable length memory and an infinite number of components. *Advances in Applied Probability*, 49, 84–107. https://arxiv.org/abs/1410.5235.

[Lambert et al., 2018] Lambert, R. C., Tuleau-Malot, C., Bessaih, T., Rivoirard, V., Bouret, Y., Leresche, N., & Reynaud-Bouret, P. (2018). Reconstructing the functional connectivity of multiple spike trains using Hawkes models. *Journal of Neuroscience Methods*, 297, 9–21.

[Ogata, 1978] Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30(2), 243–261.

[Olinde & Short, 2020] Olinde, J. & Short, M. (2020). A Self-limiting Hawkes Process: Interpretation, Estimation, and Use in Crime Modeling. *2020 IEEE International Conference on Big Data*, (pp. 3219).

[Rizoiu et al., 2017] Rizoiu, M.-A., Lee, Y., Mishra, S., & Xie, L. (2017). A Tutorial on Hawkes Processes for Events in Social Media. Preprint at http://arxiv.org/abs/1708.06401.

Processus Gaussiens

Audio Signal Denoising with Reduced Gaussian Distributions

Anis Fradi [1] & Chafik Samir [2] & Achref Ouni [2]

[1] *Université Lumière Lyon 2, Université Claude Bernard Lyon 1, ERIC, 69007, Lyon, France, anis.fradi@univ-lyon2.fr*
[2] *Laboratoire LIMOS, CNRS UMR 6158, Université Clermont Auvergne, 63170 Aubière, France*

Résumé. Les représentations probabilistes en temps-fréquence, bien qu'ayant un potentiel pour les applications de traitement du signal, en particulier pour le signal audio, ont été largement négligées en raison de leur complexité computationnelle et de la difficulté à interpréter les modèles sous-jacents. Cependant, des recherches récentes ont montré que ces représentations peuvent être modélisées à l'aide des mélanges spectraux des processus gaussiens (SM-GP), offrant ainsi une solution à ces défis et permettant une analyse probabiliste plus efficace et interprétable. Dans ce papier, nous étendons ce travail en introduisant une amélioration significative en termes d'efficacité computationnelle, de flexibilité et d'applicabilité dans des contextes réels. Notre approche exploite une technique récente d'approximation de la covariance pour accélérer l'inférence par SM-GP, tout en utilisant une méthode dans le domaine de fréquence pour l'apprentissage des hyperparamètres.

Mots-clés. processus gaussien, régression, débruitage, signal, complexité.

Abstract. Probabilistic time-frequency representations, though highly promising for signal processing applications, particularly in audio, have been largely overlooked due to their computational complexity and the difficulty in interpreting the underlying models. However, recent research has revealed that these representations can be modeled using Spectral Mixtures Gaussian Processes (SM-GP), which offers a solution to these challenges and enables more tractable and interpretable probabilistic analysis. In this paper, we extend this work by introducing a significant improvement in terms of computational efficiency, flexibility, and real-world applicability. Our approach leverages a recent covariance approximation technique to accelerate SM-GP inference, alongside a frequency-domain method for hyperparameter learning.

Keywords. Gaussian process, regression, denoising, signal, complexity.

1 Gaussian Process Regression

A GP is a probabilistic model for predicting unknown functions [6]. Defined by a collection of random variables with a joint Gaussian distribution, a GP model a distribution over functions, providing predictions with uncertainty estimates. It is fully specified by a mean function $m(t)$ and a covariance function $k(t, t')$:

$$m(t) = \mathbb{E}(f(t)); \quad k(t, t') = \mathbb{E}((f(t) - m(t))(f(t') - m(t'))).$$

Typically, $m \equiv 0$ (centered GP).

GP regression fits a function from inputs t_i to outputs y_i, accounting for prediction uncertainty using a training set of a size N: $(\mathbf{t}, \mathbf{y})$. With a GP prior and noise term $\epsilon_i \sim \mathcal{N}(0, \gamma^2)$, we have:

$$f(t) \sim \mathcal{GP}(m(t), k(t, t')) \quad \text{and} \quad y_i = f(t_i) + \epsilon_i, \quad i = 1, \ldots, N.$$

Based on the conjugate property of the Gaussian distribution, the posterior distribution over $\mathbf{f} = f(\mathbf{t})$ is also Gaussian: $p(\mathbf{f}|\mathbf{y}) = \mathcal{N}(\bar{\mathbf{f}}, \boldsymbol{\Sigma})$. From Bayes' rule, the posterior mean and covariance are:

$$\bar{\mathbf{f}} = \mathbf{K}(\mathbf{K} + \gamma^2 \mathbf{I}_N)^{-1}\mathbf{y}; \qquad \boldsymbol{\Sigma} = \mathbf{K} - \mathbf{K}(\mathbf{K} + \gamma^2 \mathbf{I}_N)^{-1}\mathbf{K}$$

where $\mathbf{K}$ is the $N \times N$ covariance matrix with entries $\mathbf{K}_{i,j} = k(t_i, t_j)$ and $\mathbf{I}_N$ refers to the $N \times N$ identity matrix. The covariance function $k(t, t')$ usually depends on a set of hyperparameters θ_k that need to be fitted from the training data.

In general, the most widely used covariances (in theory and practice) are stationary covariances (or kernels), for which $k(t, t')$ is a function of $\tau = t - t'$. For such covariances, the Bochner's theorem [1] states that they can be equivalently represented in terms of their spectral density, giving rise to the Fourier duality of covariance and spectral density:

$$k(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} S(w)e^{iw\tau}dw; \qquad S(w) = \int_{-\infty}^{+\infty} k(\tau)e^{-iw\tau}d\tau$$

where $S(.)$ is called the spectral density associated to the $\mathbb{L}^1$ kernel k.

2 Spectral Mixtures Gaussian Processes (SM-GP)

The spectral mixture kernel, as shown in [7], is defined as:

$$c(t, t') = \sum_{d=1}^{D} c^{(d)}(t, t'), \quad t, t' \in \mathbb{R}$$

where $c^{(d)}(t, t') = \cos(w_d(t - t')) \times k^{(d)}(t, t')$, with w_d interpreted as the center frequency and $k^{(d)}$ as the squared exponential (SE) kernel:

$$k^{(d)}(t, t') = \sigma_d^2 \exp\left(-\frac{(t - t')^2}{2l_d^2}\right)$$

Moreover, [7] demonstrated that any stationary kernel can be approximated with this form to arbitrary precision. The key advantage of these kernels is their balance between expressive power and modeling flexibility, while still maintaining a simple closed form that facilitates analytic inference.

Recognizing that current state-of-the-art methods for probabilistic time-frequency (TF) analysis can be seen as variations of the probabilistic phase vocoder (PPV), it follows that the SM-GP can be interpreted as probabilistic filter banks, where the length-scales l_d determine the filter bandwidths. Specifically, this connection facilitates the consideration of higher-order Markov extensions of the PPV by employing SM-GP models with smoother Matérn kernels, as opposed to the exponential kernel. The Matérn kernel is given by:

$$k^{(d)}(t, t') = \sigma_d^2 \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{|t - t'|}{l_d}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{|t - t'|}{l_d}\right)$$

where Γ is the gamma function, K_ν is the modified Bessel function of the second kind, and σ_d^2, l_d, and ν refer to the variance, length-scale, and smoothness hyperparameters, respectively. Note that the parameter ν quantifies the regularity of the GP realization.

For example, using the Matérn-$\frac{3}{2}$ ($\nu = 3/2$) and Matérn-$\frac{5}{2}$ ($\nu = 5/2$) kernels corresponds to second-order auto-regressive AR(2) and third-order AR(3) PPVs, respectively. These higher-order models are particularly useful for modeling signals with slowly varying instantaneous frequencies, which is a common characteristic in real-world signals such as audio.

Inference in these models can be performed using Kalman smoothing, with computational complexity that scales linearly with the sample size N but cubicly with the state space dimension: $O(N\eta^3)$. For a Matérn-ν covariance, the state space dimension is $\eta = (2\nu + 1)D \in \mathbb{N} \setminus \{1, 2\}$. Consequently, inference may become computationally demanding for higher-order models. Additionally, it is important to note that the SE spectral mixture kernel ($\nu \to \infty$) cannot be exactly represented within this framework.

The parameters of the SM time-frequency representation are $\theta_{\text{sm}} = (\theta_k^d, w_d)_{d=1}^D$, where θ_k^d represents the parameters of the canonical covariance $k^{(d)}(t, t')$, and w_d are the filter bank center frequencies. To learn the SM-GP hyperparameters $\theta = (\theta_{\text{sm}}, \gamma^2)$, including the noise variance, one approach is to minimize the energy function, which is equivalent to maximizing the marginal likelihood:

$$p(\mathbf{y}|\theta) = \mathcal{N}(0, \sum_{d=1}^{D} \mathbf{O}^{(d)} \odot \mathbf{K}^{(d)} + \gamma^2 \mathbf{I}_N)$$

where $\mathbf{K}_{i,j}^{(d)} = k^{(d)}(t_i, t_j)$, $\mathbf{O}_{i,j}^{(d)} = \cos(w_d(t_i - t_j))$ for $i, j = 1, \ldots, N$, and $\odot$ denotes component-wise multiplication. Optimization in the time domain is computationally expensive because it requires inverting large covariance matrices. To address this challenge, [4] proposed a frequency-domain approach by considering the Fourier transform of the marginal likelihood. Let $\tilde{\mathbf{y}} = (\tilde{y}_j)_{j=1}^N$ be the signal spectrum, and $S^{(d)}(\cdot)$ the spectral density corresponding to the covariance $k^{(d)}(\cdot)$. The spectral density associated with $c^{(d)}(\cdot)$ is $\frac{1}{2}[S^{(d)}(w - w_d) + S^{(d)}(w + w_d)]$. Therefore, the model's spectral density is:

$$\gamma_{y,i}(\theta) = \frac{1}{2} \sum_{d=1}^{D} [S^{(d)}(w_i - w_d) + S^{(d)}(w_i + w_d)] + N\gamma^2$$

where $w_i = 2\pi \frac{t_i}{N}$ is the center frequency of bin i. The log marginal likelihood in the frequency domain, also known as the "Whittle" log likelihood, is then:

$$\log p(\mathbf{y}|\theta) = -\frac{1}{2} \sum_{i=1}^{N} \left(\log(\gamma_{y,i}(\theta)) + \frac{|\tilde{y}_i|^2}{\gamma_{y,i}(\theta)} \right) - \frac{N}{2} \log(2\pi)$$

The model hyperparameters can be fitted by minimizing the dissimilarity between the model spectrum and the signal power spectrum, for example, using a gradient descent algorithm.

A key advantage of this method is its computational efficiency: the log marginal likelihood can be evaluated in $O(N + D)$ time, assuming the spectral density is known and analytic (e.g., for SE and Matérn kernels). This efficiency arises because the primary computational cost comes from inverting the vector $(\gamma_{y,1}(\theta), \ldots, \gamma_{y,N}(\theta))^T$, where each element involves summing over D terms.

However, a practical issue with the frequency-domain learning algorithm is that the signal power spectrum $|\tilde{y}_i|^2$ may exhibit large variations, potentially leading to many local optima during optimization. This issue can be mitigated by replacing the spectrum with Welch's periodogram [5], computed from multiple data segments. This produces a smoothed spectrum, which reduces the number of local optima during optimization. The effect of this smoothing diminishes as the number of data segments decreases, ultimately allowing the optimization process to more accurately fit the original spectrum.

3 Laplace Operator for Covariance Approximation

The covariance function decomposition, as stated in Mercer's theorem, asserts that a covariance $k(t,t')$ of a GP can be represented in terms of orthogonal basis functions (ϕ_l) and eigenvalues λ_l:

$$k(t,t') = \sum_{l=1}^{\infty} \lambda_l \phi_l(t)\phi_l(t') \quad t,t' \in \mathbb{R}.$$

For stationary kernels, [2] proposed approximating the covariance through an eigenfunction expansion of the Laplace operator over a compact interval $[-T,T]$, leading to:

$$k(t,t') \approx \sum_{l=1}^{\infty} S(\sqrt{\gamma_l})\psi_l(t)\psi_l(t'),$$

where $S(.)$ denotes the spectral density, and the eigenfunctions ψ_l and eigenvalues γ_l are:

$$\psi_l(t) = \frac{1}{\sqrt{T}} \sin\left(l\pi \frac{(t+T)}{2T}\right) \quad \text{and} \quad \gamma_l = \left(\frac{l\pi}{2T}\right)^2.$$

This approach allows for fast inference when the spectral density is analytically defined, as with SE and Matérn kernels:

$$S^{\text{SE}}(w) = \sigma^2\sqrt{2\pi l^2}\exp(-2\pi^2 l^2 w^2),$$

$$S^{\text{Matérn}}(w) = 2\sigma^2\pi^{1/2}\frac{\Gamma(2\nu)}{\Gamma(\nu+1/2)}\left(\frac{\sqrt{2\nu}}{l}\right)^{2\nu}\left(\left(\frac{\sqrt{2\nu}}{l}\right)^2 + w^2\right)^{-\nu-1/2}.$$

The truncated approximation of order M:

$$k_M(t,t') = \sum_{l=1}^{M} S(\sqrt{\gamma_l})\psi_l(t)\psi_l(t'),$$

converges uniformly to $k(t,t')$ as $T \to \infty$ and $M \to \infty$ [3]. The convergence rate depends on the eigenvalue decay, with smoother covariance functions generally converging faster than rougher ones, see some examples in Figure 1.

Given $\mathbf{\Psi}$ as the $N \times M$ matrix of eigenfunctions and $\mathbf{S}$ as the diagonal matrix of eigenvalues, the approximate log marginal likelihood is:

$$\log p(\mathbf{y}|\theta) \approx -\frac{1}{2}\log|\mathbf{\Psi S \Psi}^T + \gamma^2\mathbf{I}_N| - \frac{1}{2}\mathbf{y}^T(\mathbf{\Psi S \Psi}^T + \gamma^2\mathbf{I}_N)^{-1}\mathbf{y} - \frac{N}{2}\log(2\pi).$$

4 Inference in SM-GP with Reduced-Rank Covariance

We consider the SM-GP model for time-frequency representation:

$$f^{(d)}(t) \sim \mathcal{GP}(0, c^{(d)}(t,t')) \quad \text{and} \quad y_i = \sum_{d=1}^{D} f^{(d)}(t_i) + \epsilon_i, \quad i = 1,\ldots,N$$

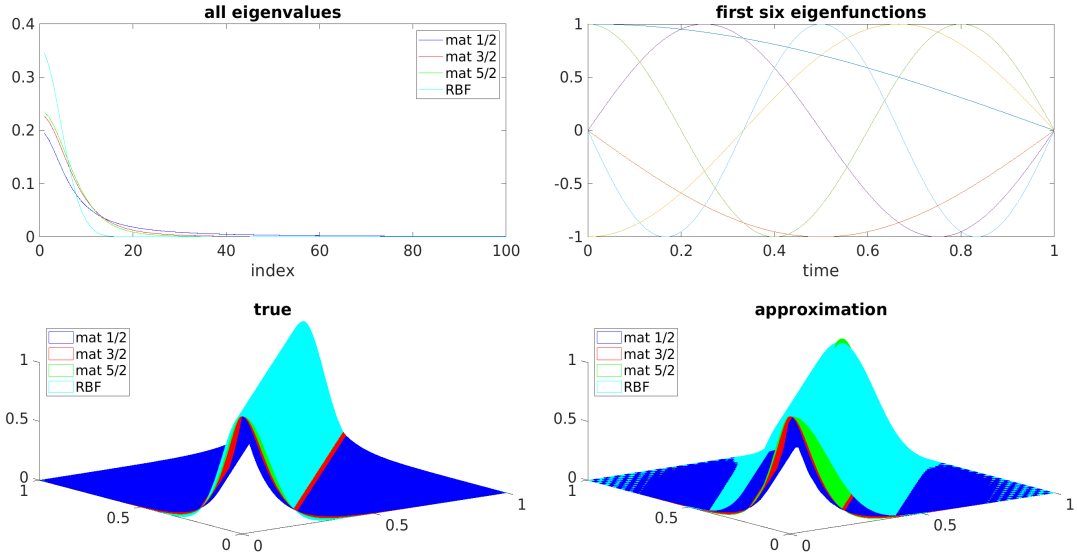


Figure 1: The eigenvalues and some eigenfunctions of the Matérn covariance function (top). An approximation of the Matérn kernel with $M = 30$ (bottom).

with the covariance for the d-th component given by:

$$c^{(d)}(t, t') = \cos(w_d(t - t')) \times k^{(d)}(t, t'),$$

where $k^{(d)}$ is a kernel with spectral density $S^{(d)}$. Assuming independence of the components, the full covariance is: $c(t, t') = \sum_{d=1}^{D} c^{(d)}(t, t')$. Let $\mathbf{f}$ denote all stacked realizations, and $\mathbf{C}^{(d)}$ the prior covariance matrix for each component. The prior and likelihood are:

$$p(\mathbf{f}) \quad = \quad \mathcal{N}(0, \mathbf{\Gamma}); \qquad p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{P}\mathbf{f}, \gamma^2 \mathbf{I}_N),$$

where $\mathbf{\Gamma}$ is a block-diagonal matrix of prior covariances and $\mathbf{P}$ selects entries from $\mathbf{f}$.

The posterior distribution of $\mathbf{f}$ is Gaussian: $p(\mathbf{f}|\mathbf{y}) = \mathcal{N}(\bar{\mathbf{f}}, \mathbf{\Sigma})$, with mean and covariance:

$$\bar{\mathbf{f}} = \mathbf{\Gamma}\mathbf{P}^T(\mathbf{P}\mathbf{\Gamma}\mathbf{P}^T + \gamma^2 \mathbf{I}_N)^{-1}\mathbf{y}; \qquad \mathbf{\Sigma} = \mathbf{\Gamma} - \mathbf{\Gamma}\mathbf{P}^T(\mathbf{P}\mathbf{\Gamma}\mathbf{P}^T + \gamma^2 \mathbf{I}_N)^{-1}\mathbf{P}\mathbf{\Gamma}.$$

The posterior mean for each component is: $\bar{\mathbf{f}}^{(d)} = \mathbf{C}^{(d)}(\mathbf{P}\mathbf{\Gamma}\mathbf{P}^T + \gamma^2 \mathbf{I}_N)^{-1}\mathbf{y}$. Using an approximation for each kernel, the covariance function is:

$$c(t, t') \approx \sum_{d=1}^{D} \cos(w_d(t - t')) \times \sum_{l=1}^{M} S^{(d)}(\sqrt{\gamma_l})\psi_l(t)\psi_l(t'),$$

see some examples in Figure 2. The posterior mean is then approximated as: $\bar{\mathbf{f}}^{(d)} = \mathbf{O}^{(d)} \odot \mathbf{\Psi}\mathbf{S}^{(d)}\mathbf{\Psi}^T\hat{\mathbf{\Sigma}}_y^{-1}\mathbf{y}$.

This approach significantly reduces computational complexity, especially when the sample size N is large. The total inference cost is $O(M^2D^2N)$, which is linear in N and quadratic in both the number of mixtures M and truncation order D. This is more efficient than both the standard SM-GP, which has cubic complexity in the sample size N, and the SM-GP based on Kalman filtering, which has cubic complexity in the state space dimension. Additionally, the proposed method can handle any stationary kernel, not just Matérn kernels.

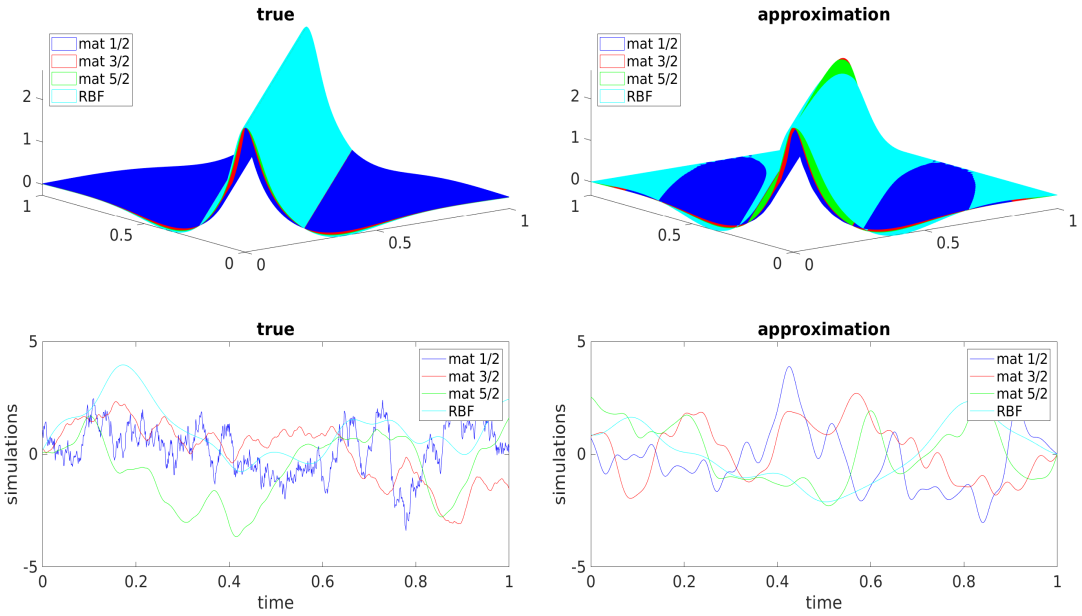


Figure 2: An approximation of a mixture of $D = 5$ Matérn kernels with $M = 30$ (top). Simulations drawn from both true SM-GP and its approximation (bottom).

References

[1] Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, page 1257–1264. MIT Press, 2005.

[2] Arno Solin and Simo Särkkä. Hilbert space methods for reduced-rank Gaussian process regression. *Statistics and Computing*, 30(2):419–446, 2020.

[3] R. Takhanov. On the speed of uniform convergence in Mercer's theorem. *Journal of Mathematical Analysis and Applications*, 518:126718, 2023.

[4] Richard E. Turner and Maneesh Sahani. Time-frequency analysis as probabilistic inference. *IEEE Transactions on Signal Processing*, 62(23):6171–6183, 2014.

[5] Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15:70–73, 1967.

[6] Christopher K. I. Williams and David Barber. Bayesian classification with Gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(12):1342–1351, 1998.

[7] Andrew Gordon Wilson and Ryan Prescott Adams. Gaussian process kernels for pattern discovery and extrapolation. In *ICML (3)*, pages 1067–1075, Atlanta, Georgia, USA, 2013. Proceedings of Machine Learning Research (PMLR).

Wavelet Gaussian process regression for functional output

Baptiste Kerleguer [1]

[1] *CEA, DAM, DIF, F-91297, Arpajon, France & baptiste.kerleguer@cea.fr*

Résumé. La régression par processus gaussien est largement utilisée pour émuler la sortie d'un code coûteux. Nous nous intéressons, dans ce travail, à des codes dont la sortie est une fonction du temps. Pour réaliser des métamodèles de ces codes, il est usuel de réduire la dimension puis de réaliser une régression dans l'espace latent, à l'aide de la régression par processus gaussien par exemple. Après avoir mis en lumière les problèmes de l'hypothèse d'indépendance des paramètres dans l'espace latent sur les sorties, ce travail introduit les processus gaussiens d'ondelette afin d'éviter de réaliser une régression par processus gaussien en grande dimension. Cette méthode permet de conserver les propriétés intéressantes de la régression par processus gaussiens tout en facilitant l'implémentation. Les processus ainsi créés sont comparés avec les méthodes usuelles sur un exemple analytique.

Mots-clés. Processus Gaussien, modèle de subtitution, ondelette

Abstract. Gaussian process regression is widely used to emulate the output of expensive code. In this work, we are interested in codes whose output is a function of time. To produce surrogate model of these codes, it is usual to reduce the dimension and then perform a regression in latent space, using Gaussian process regression for example. After highlighting the problems of the assumption of independence of the parameters in latent space on the outputs, this work introduces wavelet Gaussian processes in order to avoid performing a large dimension Gaussian process regression. This method preserves the interesting properties of Gaussian process regression, making it easier to implement. The processes thus created are compared with the usual methods on an analytical example.

Keywords. Gaussian process, surrogate model, wavelet

1 Introduction

Advances in scientific modeling have led to the development of more complex and computationally expensive codes. It has therefore become necessary to use surrogate models, constructed from the outputs of these codes, in order to study the uncertainties of these codes. One method, also known as Kriging, widely used in the uncertainty quantification community to produce surrogate models is Gaussian process (GP) regression, see Gramacy (2022) and Williams and Rasmussen (2006).

As the complexity of scientific modeling has increased, model outputs have become more complex. One of the difficulties is the high dimension of the outputs. We are interested in

codes with high-dimensional outputs whose outputs are time-dependent functions. When they are sampled, they are also known as time series. We assume that the sampling is done on a fine, regular grid, so the output is high-dimensional. It is assumed that time series are available at all times for each call to the code.

When the regression problem has high-dimensional outputs the natural solution is to reduce the dimension of the outputs to return to the previous case. The difficulty is therefore divided into two parts: dimension reduction and Gaussian process regression. Usually, these two parts are treated one after the other, with a certain degree of decoupling. In other words, we can change the dimension reduction method without changing the model in the latent space.

The problem we are interested in is the following: let us consider a computational code of which outputs are of the form $z(\mathbf{x}, t)$ with $\mathbf{x} \in \mathbb{R}$ and $t \in [0, 1]$. In the following the stochastic process emulating the code in $Z(\mathbf{x}, t)$. We know this code in a limited number of $\mathbf{x}$ but for t discretized on a regular grid. We try to reduce the output $z(\mathbf{x}, t)$ by r coefficients $a_1(\mathbf{x}), \cdots, a_r(\mathbf{x})$. For that we use a dimension reduction noted $\mathcal{F}$. The principle of dimension reduction is illustrated in Figure 1. For the regression, the objective is to predict in the space of $z(\mathbf{x}, t)$, it is thus necessary to be able to return to the output space of the code. This is how we look for a pseudo-inverse of $\mathcal{F}$ denoted $\mathcal{F}^{-1}$. Thus, we obtain an approximation $\hat{z}(\mathbf{x}, t)$ of $z(\mathbf{x}, t)$. Our objective is to have r as small as possible and to have $\hat{z}(\mathbf{x}, t)$ as close as possible to $z(\mathbf{x}, t)$. In order to solve the regression problem, GP regression will be used on the coefficients $a_1(\mathbf{x}), \cdots, a_r(\mathbf{x})$. The space of these coefficients is called the latent space.

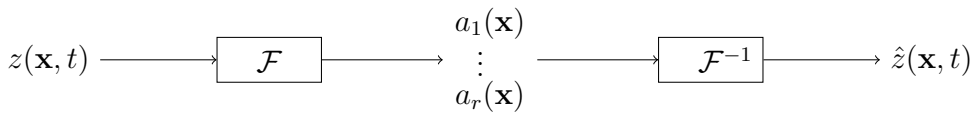


Figure 1: Illustration of the dimension reduction for time series outputs. $\mathcal{F}$ is the dimension reduction function and $\mathcal{F}^{-1}$ its pseudo inverse.

A survey of surrogate model methods for high-dimensional outputs with uncertainty quantification is available in Kontolati et al. (2022).

The method proposed in this work is called wavelet Gaussian process (WGP) regression. WGP regression uses a wavelet transform for the data to reduce the complexity of the computation conditioning. Wavelet are known to be a space representation in many fields. Moreover, Fourier transforms and wavelet transforms are commonly used to compress information from natural signals. These methods allow working in known projection spaces, which have the advantage of concentrating the information on a few coefficients. Moreover, these methods are invertible, which is essential for regression. The output signal will therefore be decomposed on a wavelet basis. The method presented is similar to the one presented in Perrin et al. (2021) but with a covariance structure that takes into account the wavelet decomposition.

In the following section, we present the wavelet Gaussian process. This result will be compared with a Gaussian process with independent coefficients in the latent space. Then

in the third section, using an example of an analytical function, we will compare different methods.

2 Wavelet functional Gaussian process regression

The study of stochastic processes, and Gaussian processes in particular, in the wavelet framework is not new. The properties of stationary Gaussian processes have long been known, as presented in Timmer and Koenig (1995). There is also theoretical work carried out on time-series, such as that presented in Moulines et al. (2008). Work has also been done on stochastic processes and the covariance function, such as in Morel (2022).

Let's use the definition of the wavelet family $\{\psi_{j,k}, j \in \mathbb{Z}, k \in \mathbb{Z}\}$ and $\{\phi_{j,k}, j \in \mathbb{Z}, k \in \mathbb{Z}\}$ of translated and dilated functions: $\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi\left(2^{-j}t - k\right)$ and $\phi_{j,k}(t) = 2^{-\frac{j}{2}}\phi\left(2^{-j}t - k\right)$. For any square integrable function $f : \mathbb{R} \mapsto \mathbb{R}$, the wavelet coefficients $W^f_{j,k}$ are defined by:

$$W^f_{j,k} = \int f(t)\psi_{j,k}(t)\mathrm{d}t, \tag{1}$$

with $j \geq 0$ and $k \in \mathbb{Z}$. The definition set of points (j,k) is I. From the expression of $\mathbf{W}^f_{j,k}$ the parameters j and k are in I, defined by:

$$I = \left\{j, k | j = 1, \cdots, M \ \ \& \ \ k = 0, \cdots, 2^{M-j} - 1\right\}. \tag{2}$$

I is a vector of $2^M - 1$ elements, and with the addition of a^M, $\left(a^M, \mathbf{W}^f_M, \cdots, \mathbf{W}^f_1\right)$ is 2^M-dimensional.

Definition 1 *(Wavelet Gaussian process) Let ψ and ϕ be respectively a wavelet function and the scale function. Let Z be a Gaussian process with tensorized covariance $k = k_x F$. The associated Gaussian process A is introduced such that the process Z can be expressed as:*

$$Z(\mathbf{x}, t) = A(\mathbf{x}, 0, 0)\phi(t) + \sum_{j,k \in I} A(\mathbf{x}, j, k)\psi_{j,k}(t), \tag{3}$$

where I is defined in eq.(2), $I = \bigcup_{M \to \infty} \left\{j, k | j = 1, \cdots, M \ \ \& \ \ k = 0, \cdots, 2^{M-j} - 1\right\}$. A is called the (ϕ,ψ) WGP of Z.

The formula for the covariance of the Gaussian process A can be calculated. The general case has been treated as well as the special case of the Haar transform and the Matérn kernel $5/2$. This makes it possible to perform Gaussian process regression in wavelet space. Since the discrete wavelet transform is a linear transformation, it is straightforward to compute the covariance matrix in wavelet space as a function of the covariance matrix in time space.

Let A be the WGP associated to Z. As the outputs of the code are time-series, N_t data points $(\mathbf{x}, t_u)_{u=1}^{N_t}$ are available for all points in the training set $\mathbf{x}$. Without lost of generality, N_t is assumed to be a power of 2, $N_t = 2^M$ with $M \in \mathbb{N}$. At the observed data $z(\mathbf{x}, t_u)$ of the

GP $Z(\mathbf{x}, t_u)$, $a(\mathbf{x}, j, k)$ is associated with $(j, k) \in I$. The WGP $A(\mathbf{x}, j, k)$ can be constructed with $(j, k) \in I$ because $\text{card}(I) = N_t$. The equivalent of Gaussian process regression for A is $A(\mathbf{x}, j, k) \sim \mathcal{GP}\left(0, k(\mathbf{x}, \mathbf{x}')k_{\text{WGP}}(j, k, j', k')\right)$, with k a kernel on x. The input observation set is denoted $\mathcal{D}_{x,t} = \mathcal{D}_x \times \mathcal{D}_t$ with $\mathcal{D}_x = \{x_i\}_{i=1,\dots,N_x}$ and $\mathcal{D}_t = \{(j, k) \in I\}$. The posterior of A knowing the observations $a^{\mathbf{N_x}} = \{a(\mathbf{x}_i, j, k)\}_{i=1,\dots,N_x}^{(j,k)\in I}$ and the hyperparameters θ is:

$$A(\mathbf{x}, j, k)|\theta, A(\mathcal{D}_{x,t}) = a^{\mathbf{N_x}} \sim \mathcal{GP}\left(\mu^\star(\mathbf{x}, j, k), c^\star(\mathbf{x}, j, k)\right), \tag{4}$$

with :

$$\left\{ \begin{array}{lcl} \mu^\star(\mathbf{x}, j, k) & = & r(\mathbf{x})^T \mathbf{R}_{x,\text{WGP}}^{-1} a^{\mathbf{N_x}} \\ c^\star(\mathbf{x}, j, k) & = & k_x(\mathbf{x}, \mathbf{x}')k_{\text{WGP}}(j, k, j', k') - r(\mathbf{x})^T \mathbf{R}_{x,\text{WGP}}^{-1} r(\mathbf{x}') \end{array} \right. , \tag{5}$$

with $r(\mathbf{x}') = [k(\mathcal{D}_x, \mathbf{x}')k_{\text{WGP}}(\mathcal{D}_t, j', k')]$ and $\mathbf{R}_{x,\text{WGP}} = [k(\mathcal{D}_x, \mathcal{D}_x)k_{\text{WGP}}(\mathcal{D}_t, \mathcal{D}_t)]$. The regression is proceeded with $N_t \times N_x$ learning points $(\mathbf{x}, j, k)$. The expressions of the mean and variance of the prediction, in equation 5, limits the number of learning points. This is due to the needed inversion of $\mathbf{R}_{x,\text{WGP}}$. Let N_{Krig} be the maximum number of points that can be considered for GP regression. Because $N_t \times N_x > N_{\text{Krig}}$ the selection of point is the key to the regression.

The kernels of the Gaussian process regression k_x and k_t depend on hyperparameters. The inventory of hyperparameters gives us a coefficient of standard deviation σ and a length l_x. We optimize the hyperparameters in wavelet space and achieve a prediction on $Z(\mathbf{x}, t)$. This method can therefore be compared with other high-dimensional Gaussian process regression methods. This is done in the next section.

3 Example

As mentioned in Raissi (2016), the approximation of a discontinuous function using GP regression is challenging due to the continuous kernel employed. To exploit the full potential of the Gaussian process regression method, it is useful to use a discontinuous function. Therefore, a simple analytical function with a jump in t that depends on the parameter in x is proposed in the following equation:

$$f(x, t) = \sin\left(4\pi(\frac{x}{4} + 1)t\,\text{sgn}(x - t) - \frac{x}{10}\right), \tag{6}$$

with sgn the sign function. The function is evaluated in a set of N_x point elements x and for each point of the temporal grid. The time grid is composed of N_t points regularly spaced between 0 and 1. Here $N_x = 50$ and $N_t = 64$. The conditions for computing R^2 are those set out in the section introduction. The PCA method used assumes independence between coefficients and takes the 5 most important coefficients.

For this function, since the temporal correlation depends on x, the proposed methods should have average performance, since they all implicitly or explicitly assume that the correlation in t and x are independent. Indeed, Figure 2 shows that the R2 value for all the methods is no better than 0.9. However, only the WGP regression performs at this level. Moreover, the interquantile range of the errors is much smaller for the WGP regression.

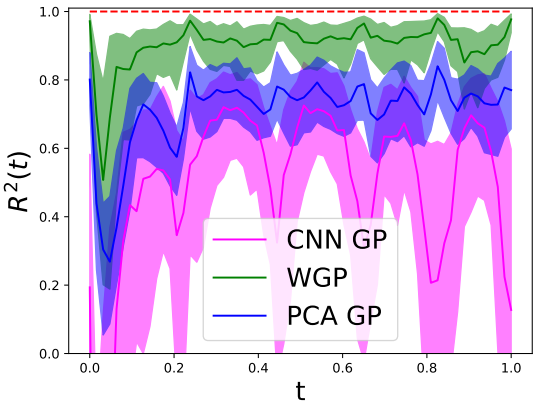


Figure 2: Error for all the methods depending on the time. The intervals are between .25 and .75 quantiles

It is therefore possible to say that the method is relevant in particular for the case of regression of temporally discontinuous functions.

4 Conclusion

This work introduces a wavelet Gaussian process (WGP) regression that uses the wavelet transform to simplify the calculation of the GP regression. This method has been shown to be theoretically equivalent, but with a well-conditioned covariance matrix. For time series, as shown in the literature, the large dimension of the output makes it impossible to compute with GP regression. This problem is less pronounced with WGP regression. The performance of WGP regression demonstrated against the neural network and singular value decomposition are fine for estimating a time series as an output in an example. It is also possible to use WGP to represent the physical constraints of the model, which opens up new possibilities for this method. One of the main limitations is that all coefficients are used for the WGP regression. If the dimension increases, it will not be possible to proceed with GP regression. This problem has been discussed, but no methods were proposed to generally solve this issue. A promising direction for future work can be to use universal GP regression.

Bibliographie

Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*, CRC press, Boca Raton.

Kontolati, K., Loukrezis, D., Giovanis, D. G., Vandanapu, L., & Shields, M. D. (2022). A survey of unsupervised learning methods for high-dimensional uncertainty quantification in black-box-type problems, *Journal of Computational Physics*, 464, 111313.

Morel, R., Rochette, G., Leonarduzzi, R., Bouchaud, J. P., & Mallat, S. (2023). Scale dependencies and self-similar models with wavelet scattering spectra, *Available at SSRN* 4516767.

Moulines, E., Roueff, F., and Taqqu, M. S. (2008). A wavelet whittle estimator of the memory parameter of a nonstationary gaussian time series. *The Annals of Statistics*, 36(4):1925–1956.

Perrin, T. V. E., Roustant, O., Rohmer, J., Alata, O., Naulin, J. P., Idier, D., Pedreros, R., Moncoulon, D. & Tinard, P. (2021). Functional principal component analysis for global sensitivity analysis of model with spatial output. *Reliability Engineering & System Safety*, 211, 107522.

Raissi, M. and Karniadakis, G. (2016). Deep multi-fidelity gaussian processes. *arXiv preprint* arXiv:1604.07484.

Timmer, J. and Koenig, M. (1995). On generating power law noise. *Astronomy and Astrophysics*, v. 300, p. 707, 300:707.

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, Cambridge, MA: MIT press.

Gaussian process regression and active subspaces recipe for parsimonious metamodeling

Amandine Marrel [1,3] & Bertrand Iooss [2,4]

[1] *CEA, DES, IRESNE, DER, Cadarache, France, amandine.marrel@cea.fr*
[2] *EDF R&D, Chatou, and SINCLAIR AI Lab., Saclay, France, bertrand.iooss@edf.fr*
[3] *LMA Université d'Avignon, EA 2151, 84029 Avignon, France*
[4] *Institut de Mathématiques de Toulouse, 31062, Toulouse, France*

Résumé. Dans le domaine de la modélisation et de la prévision de phénomènes physiques complexes, les simulateurs numériques jouent un rôle crucial. Cependant, l'incertitude sur certains paramètres d'entrée impacte la précision des prédictions et le temps d'exécution CPU peut restreindre le nombre de simulations possibles, compliquant des tâches telles que la quantification des incertitudes ou la calibration de paramètres. Pour pallier cela, la sortie du simulateur peut être approchée par un métamodèle, par exemple la régression par processus gaussien (RPG) qui offre certains avantages. Afin de déployer plus efficacement la RPG sur des modèles "boîte noire" en grande dimension, nous proposons ici une méthodologie combinant un criblage et une réduction de dimension via des sous-espaces actifs. Un criblage des entrées significatives est d'abord réalisé à l'aide de tests d'indépendance basés sur le critère d'indépendance de Hilbert-Schmidt (HSIC). Une première RGP peut alors être construite et permet d'estimer les sous-espaces actifs. Des heuristiques permettent ensuite de sélectionner un nombre limité de directions actives à conserver. Enfin, une nouvelle RPG est construite en utilisant uniquement les directions actives sélectionnées. La méthodologie est testée sur une collection de fonctions analytiques de dimensions variées, démontrant sa capacité à trouver un compromis optimal entre réduction de dimension et prédictivité de la RPG.

Mots-clés. Régression par processus gaussien, Sous-espaces actifs, Apprentissage, automatique, Criblage, Incertitude.

Abstract. In the field of modeling and predicting complex physical phenomena, numerical simulators play a crucial role. However, uncertainty in certain input parameters can limit the accuracy of predictions, and CPU execution time can restrict the number of possible simulations, complicating tasks such as uncertainty quantification or parameter calibration. To address this, the simulator output can be approximated by a metamodel, as the Gaussian Process Regression (GPR) which has many advantages. To more efficiently deploy GPR on high-dimensional "black-box" models, we propose a methodology combining screening and dimension reduction via active subspaces. Significant input variables are firstly screened using independence tests based on the Hilbert-Schmidt Independence Criterion (HSIC). This reduction enables the construction of a GPR, from which the active subspaces can be estimated. Heuristics are then used to select a limited number of active directions to retain. Finally, a new GPR is constructed using only the selected active directions. This methodology is tested on a collection of analytical functions of varying dimensions, demonstrating its ability to find an optimal compromise between dimensionality reduction and GPR predictivity.

Keywords. Gaussian process, Active Subspaces, Machine learning, Screening, Uncertainty.

1 Introduction

In risk assessment, numerical simulators are widely used to model and predict various phenomena. However, information on some input parameters, related to physical or numerical modeling, can be limited or uncertain due a lack of knowledge about the underlying physics, partial input characterization, or scenario choices. The uncertainty quantification (UQ) in simulation results is then crucial (Smith, 2014). This typically involves a probabilistic framework where input uncertainties are modeled by probability distributions based on data, expert opinions, or bibliographic sources. UQ process mainly relies on Monte Carlo techniques, performing simulations with inputs drawn from these distributions. Estimators of target quantities (probabilities, quantiles, optimal values) are then computed from the simulations.

A key issue with numerical models, especially high-fidelity models, is their time-consuming nature, limiting the number of possible simulations. This is especially problematic for complex objectives like inverse or optimization problems, which require an extensive exploration of the input space. To address this, surrogate models (or metamodels) are used to approximate the simulator output by an inexpensive mathematical model, such as polynomials, random forests, neural networks, etc. (Fang et al., 2006), essentially any machine learning technique. Built from simulation data, metamodels must accurately represent code outputs while offering good prediction capabilities. It can then serve multiple purposes, such as sensitivity analysis, uncertainty propagation, optimization, or calibration. Gaussian process regression (GPR) is a very popular metamodel due to its flexibility and appealing UQ properties (Rasmussen and Williams, 2006). GPR provides a predictive distribution for code output, enabling Bayesian optimization strategies to be deployed that balance exploration and exploitation. This distribution also yield more reliable results when searching for inputs' values that penalize the output variables or estimating conditional probabilities.

However, a large number of inputs poses challenges in GPR metamodeling, affecting both the GPR estimation and its use in Bayesian frameworks. Sensitivity analysis techniques, like the Hilbert-Schmidt independence criterion (HSIC) and the associated statistical independence tests (Gretton et al., 2012, Marrel et al., 2022), can be used to identify significantly influential inputs. This initial screening most often reduces the number of variables to fewer than 20, making reliable GPR estimation feasible with an efficient algorithm (see, e.g., Marrel et al., 2024). Nevertheless, it may be interesting to go beyond this first reduction, especially for GPR-based Bayesian inference in UQ. Obtaining a GPR defined in a smaller input space can be achieved through a low-dimensional embedding (Binois and Wycoff, 2022). Linear embedding using the active subspaces (AS) introduced by Constantine (2015) is of great interest as it identifies the directions with the most variation in the input space.

In this work, we start from the idea of Binois and Picheny (2024) who project high-dimensional inputs onto the active subspaces to define new input variables, simplifying GPR modeling and its use. We propose a methodology for meaningfully combining GPR with AS reduction for applications where gradient is not available (AS computation involves decomposing a matrix derived from the gradients of the model outputs). It aims at optimizing the trade-off between dimension reduction and metamodel predictive accuracy by selecting the optimal number of active directions via relevant heuristics.

Section 2 provides reminders on GPR and AS. Section 3 explained the proposed methodology while Section 4 gives the results of a benchmark on a collection of analytical functions of varying dimensions to evaluate the performance of the method. Section 5 concludes.

2 Reminders on Gaussian process regression and active subspaces

The numerical model (computer code or simulator) is represented by the following input-output relationship:

$$\mathcal{M} : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathcal{Y} \\ \mathbf{X} & \longmapsto & Y = \mathcal{M}(\mathbf{X}) \end{array} \right. \tag{1}$$

where $\mathbf{X} = (X_1, \ldots, X_d)^\top$ are the d inputs belonging to some measurable space $\mathcal{X} \subset \mathbb{R}^d$. In the context of given-data and data-driven GPR metamodeling, we only have a n-size learning sample of inputs and outputs denoted by $(\mathbf{X_s}, \mathbf{Y}(\mathbf{X_s}))$ where $\mathbf{X_s} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ (with $\mathbf{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_d^{(i)}\right)$) denotes the matrix of n-size sample locations, and $\mathbf{Y_s} = \{y^{(1)}, \ldots, y^{(n)}\}$ denotes the corresponding outputs observations with $y^{(i)} = \mathcal{M}(\mathbf{x}^{(i)})$.

2.1 Gaussian process metamodel and associated parametric choices

In the GP regression (Rasmussen and Williams, 2006), Y is considered as a realization of a Gaussian stochastic process:

$$Y(\cdot) \sim \mathcal{N}\left(m(\cdot), k(\cdot,\cdot)\right), \tag{2}$$

where $m(\cdot)$ is the mean function and $k(\cdot,\cdot)$ is the covariance kernel function. The predictive GP distribution is then the GP conditioned by the learning sample $(\mathbf{X_s}, \mathbf{Y_s})$:

$$Y(\cdot) \mid \mathbf{Y_s} \sim \mathcal{N}\left(\hat{y}(\cdot), \hat{c}(\cdot,\cdot)\right), \tag{3}$$

with

$$\hat{y}\left(\mathbf{x}\right) = \mathbb{E}[Y(\mathbf{x})|\mathbf{Y_s}] = m(\mathbf{x}) + \mathbf{k}(\mathbf{x}, \mathbf{X_s})^\top \mathbf{K}^{-1}\Big(\mathbf{Y_s} - \mathbf{m}(\mathbf{X_s})\Big), \tag{4}$$

and

$$\hat{c}(\mathbf{x}, \tilde{\mathbf{x}}) = \mathbb{COV}[Y(\mathbf{x}), Y(\tilde{\mathbf{x}})|\mathbf{Y_s}] = k(\mathbf{x}, \tilde{\mathbf{x}}) - \mathbf{k}(\mathbf{x}, \mathbf{X_s})^\top \mathbf{K}^{-1} \mathbf{k}(\tilde{\mathbf{x}}, \mathbf{X_s}), \tag{5}$$

where $\mathbf{m}(\mathbf{X_s}) = (m(\mathbf{x}^{(i)}))_{1 \leq i \leq n} \in \mathbb{R}^n$, $\mathbf{K} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}))_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}$ and $\mathbf{k}(\mathbf{x}, \mathbf{X_s}) = (k(\mathbf{x}, \mathbf{x}^{(i)}))_{1 \leq i \leq n} \in \mathbb{R}^n$. The efficiency of GPR strongly depends on the specifications of $m(\cdot)$ and $k(\cdot,\cdot)$ functions. In practice, this consists of making parametric choices among a panel of usual functions. Most often, a constant or a one-degree polynomial is chosen for $m(\cdot)$ and a tensorized covariance built-upon d 1-D Matérn covariance functions is considered:

$$k_{\sigma,\boldsymbol{\theta}}(\mathbf{x}, \tilde{\mathbf{x}}) = \sigma^2 \prod_{i=1}^{d} k_{\theta_i}(x_i - \tilde{x}_i). \tag{6}$$

where σ^2 is the variance parameter and $\theta_i \in \mathbb{R}^+$ are correlation hyperparameters.

Estimating the GPR therefore boils down to estimating its parameters from the learning sample. Three main estimation procedures are used: minimization of prediction error calculated by cross-validation, maximization of likelihood, and Bayesian approach. Once estimated, the predictive capabilities of the GPR need to be validated. Common validation criteria include the predictivity coefficient Q^2 and the IAEα criterion, to assess the accuracy of the GPR predictor $\hat{y}(\mathbf{x})$ and the reliability of GPR prediction intervals, respectively (see, e.g., Marrel and Iooss, 2024a). In practice, in the small-data application context, validation criteria are computed by cross-validation (CV).

2.2 Active Subspaces for Linear Embedding

Active subspaces (AS), as defined by Constantine (2015), are a dimensionality reduction technique that identifies the most significant directions of variation in the input space of a function through gradient analysis. More precisely, searching for a linear embedding is equivalent to finding the matrix $\mathbf{A}$ such that $\mathcal{M}(\mathbf{x}) \approx f(\mathbf{A}^\top \mathbf{x})$, where $\mathbf{A}$ is a $d \times q$ matrix, with q being the reduced dimension ($q \ll d$). In the AS approach, $\mathbf{A}$ is defined by the largest q eigenvectors of the covariance matrix $\mathbf{C} \in \mathbb{R}^{d \times d}$:

$$C = \int_{\mathcal{X}} \nabla \mathcal{M}(\mathbf{x}) \nabla \mathcal{M}(\mathbf{x})^\top \lambda(d\mathbf{x})$$

with $\nabla \mathcal{M}(\mathbf{x}) = \left(\frac{\partial \mathcal{M}(\mathbf{x})}{\partial x_1}, \frac{\partial \mathcal{M}(\mathbf{x})}{\partial x_2}, \ldots, \frac{\partial \mathcal{M}(\mathbf{x})}{\partial x_d} \right)$ and λ is a measure on the input space $\mathcal{X}$. The covariance matrix $\mathbf{C}$ is then decomposed using eigenvalue decomposition: $\mathbf{C} = V \Lambda V^\top$ where V is the matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues. High-dimensional input variables $\mathbf{x}$ are projected onto the active subspaces defined by the subset of eigenvectors $V_1 \in \mathbb{R}^{d \times q}$ corresponding to the largest q eigenvalues. The new input variables are given by: $\mathbf{z} = V_1^\top \mathbf{x}$, where $\mathbf{z} \in \mathbb{R}^q$ is the lower-dimensional representation of the input space.

So, in practice, the estimation of active subspaces requires access to the derivatives of the model output, which are often not available in the case of black-box models. To circumvent this problem, we propose to estimate the $\mathbf{C}$ matrix (and therefore the AS) from an initial high-dimensional GPR built on the set of input variables $\mathbf{x}$. It remains the question of how many components to keep: the challenge is to find the value of q that offers the best compromise between preserving GPR predictivity and maximizing dimension reduction.

3 Proposed methodology

To combine GPR with AS for large-dimension and gradient-free problems, we propose the following methodology in four key steps:

1. **Initial Screening built-upon dependence measures:** perform an initial screening of the input variables using HSIC-based dependence measures and associated independence tests, to retain only the significantly influential variables. This typically reduces the number of input variables to less than 20, making GPR estimation tractable.

2. **Building an initial high-dimensional GPR:** estimate the hyperparameters of a first GPR on the reduced set of input variables (from Step 1), using the ICSCREAM method (Marrel et al., 2022) and efficient estimation algorithm (Marrel and Iooss, 2024b).

3. **Estimation of AS and optimal selection:** estimate the active subspaces using the high-dimensional GPR from Step 2. To determine q the optimal number of active directions to keep, they are progressively included in the GPR: at each iteration, a GPR is built with the first j active directions as inputs. Several strategies are proposed for choosing q: the value of j that maximizes CV-Q^2 (Q^2 estimated by CV) or as soon as a stagnation of CV-Q^2 is observed (the improvement of CV-Q^2 becomes negligible).

4. **Construction of a low-dimensional GPR:** build a new GPR using only the selected AS of Step 3 as input variables. More computationally efficient, it is hoped to retain or better exceed the predictive accuracy of the high-dimensional GPR.

4 Benchmark and results

This methodology has been evaluated through a benchmark on analytical functions, for different input dimensions and learning sample sizes, and compared with a dimension reduction only based on HSIC screening. For each test case, three metrics are computed for GPR of Step 4 w.r.t. the high-dimensional GPR of Step 2: the variations of Q^2 and IAEα (noted ΔQ^2 and ΔIAEα) estimated on a large test sample, and the dimension reduction rate. Several GPR covariance models are also tested (exponential, Matérn 5/2 and Gaussian covariance).

The results, illustrated by Figure 1, show that the choice of q by maximizing the CV-Q^2 maintains the predictivity of the reduced GPR at a level very close to that of the high-dimensional GPR, while reducing the dimension by an average of 25%. The choice by CV-Q^2 stagnation (under the additional constraint that CV-Q^2 be no more than 5% lower than CV-Q^2 of the high-dimensional GPR) makes it possible (on average) to double the dimension reduction, but at the cost of a slightly larger and more variable Q^2 drop (Q^2 on test sample).

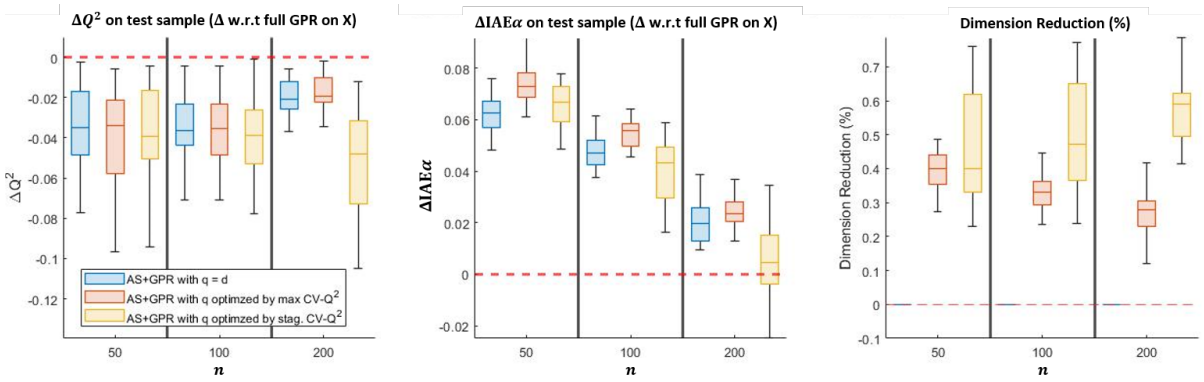


Figure 1: Results for randomized Becker's function (Marrel and Iooss, 2024b) with $d = 20$ and 30% interaction effect (boxplot over randomized functions and random learning samples).

Furthermore, it appears that the method is robust, in the sense that the reduced GPR retains the same performance as the high-dimensional GPR, even for small sample sizes. In fact, the hey point is to have a first high-dimensional GPR with a CV-Q^2 between 0.75 and 0.85 to have a reliable estimate of the active subspaces, or at least to identify a reduced space that contains all the necessary information of the inputs/output relationship. With regard to the IAEα criterion reflecting the quality of prediction intervals, a slight decrease is observed on average over all tests, limited to 0.05, which remains acceptable, but which deserves to be investigated and explained.

5 Conclusion

The combination of HSIC-based screening and Gaussian Process Regression (GPR) with Active Subspace (AS) reduction provides an effective methodology for handling high-dimensional, gradient-free problems. Sequencing HSIC-based variable selection and GPR building on the reduced set of variables, we can reliably estimate AS. We can then build a new GPR on a more reduced space derived from AS-based linear embedding of the initial inputs. This approach thus achieves a good compromise between dimension reduction and GPR prediction accuracy. It therefore paves the way for a more optimal and less costly use of GPR for uncertainty quantification, calibration or optimization tasks in the Bayesian framework.

References

Binois, M. et Picheny, V. (2024), Combining additivity and active subspaces for high-dimensional gaussian process modeling, arXiv:2402.03809.

Binois, M. et Wycoff, N. (2022), A survey on high-dimensional gaussian process modeling with application to bayesian optimization, *ACM Trans. Evol. Learn. Optim.*, 2.

Constantine, P. G. (2015), *Active Subspaces*, SIAM, Philadelphia, PA, USA.

Fang, K.-T., Li, R. et Sudjianto, A. (2006), *Design and Modeling for Computer Experiments*, Chapman & Hall/CRC.

Gretton, A., Borgwardt, K.-M., Rasch, J.-M., Schölkopf, B. and Smola, A. (2012), A kernel two-sample test, *Journal of Machine Learning Research*, 13, pp. 723–773.

Marrel, A., Iooss, B. and Chabridon, V. (2022), The ICSCREAM methodology: Identification of penalizing configurations in computer experiments using screening and metamodel – Applications in thermal-hydraulics, *Nuclear Science and Engineering*, 196, pp. 301–321.

Marrel, A. and Iooss, B. (2024a), Probabilistic surrogate modeling by Gaussian process: A review on recent insights in estimation and validation, *Reliab. Eng. Syst. Saf.*, 247, 110094.

Marrel, A. and Iooss, B. (2024b), Probabilistic surrogate modeling by Gaussian process: A new estimation algorithm for more robust prediction, *Reliab. Eng. Syst. Saf.*, 247, 110120.

Rasmussen, C. and Williams, C. (2006), *Gaussian Processes for Machine Learning*, MIT Press.

Smith, R. (2024), *Uncertainty quantification, Second Edition*, SIAM.

Régression avancée

Regularisation Ridge de la régression PLS

Luca Castelli[1] & Irène Gannaz[2] & Clément Marteau[1]

[1] Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, F-69622 Villeurbanne, France, castelli@math.univ-lyon1.fr marteau@math.univ-lyon1.fr
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP, 38000 Grenoble, France, irene.gannaz@grenoble-inp.fr

Résumé. L'algorithme de la régression PLS (*Partial Least Squares*) permet d'estimer le paramètre de la régression linéaire $Y = X\beta + \varepsilon$ dans un contexte de haute dimension. Nous proposons une régularisation Ridge de l'estimateur PLS, basé sur l'écriture dans un espace de Krylov. Nous fournissons des bornes au risque de prédiction. Nos bornes sont non asymptotiques et sont exprimées en fonction du nombre d'observations, du niveau de bruit, des propriétés de la matrice de covariables X et du nombre de composantes PLS considérées. En particulier, nous présentons certains scénarios dans lesquels la variabilité de l'estimateur PLS est importante et nous prouvons que la régularisation de Ridge permet alors d'améliorer les performances de l'estimateur. Ces résultats théoriques sont illustrés par des simulations numériques.

Mots-clés. Partial least squares, Réduction de dimension, Régression, Régularisation Ridge

Abstract. The Partial Least Squares (PLS) regression algorithm allows to estimate the linear regression parameter in the model $Y = X\beta + \varepsilon$ in a high-dimensional context. We propose a Ridge regularization of the PLS estimator, based on the writing in a Krylov space. We provide bounds on the prediction risk. Our bounds are non-asymptotic and are expressed as a function of the number of observations, the noise level, the properties of the covariate matrix X and the number of PLS components considered. In particular, we present some scenarios in which the variability of the PLS estimator is important and we prove that the Ridge regularization then improves the performances of the estimator. These theoretical results are illustrated by numerical simulations.

Keywords. Partial least squares, dimension reduction, regression, Ridge regularization

1 Introduction

Nous nous intéressons au modèle linéaire classique dans un contexte de grande dimension. Nous observons un échantillon de taille n, (X_i, Y_i), $i = 1, \ldots, n$, où les $Y_i \in \mathbb{R}$ sont les variables de sortie et les $X_i \in \mathbb{R}^p$ sont les covariables p-dimensionnelles. Nous considérons une relation linéaire au sein de chaque couple (X_i, Y_i), représentée par l'équation :

$$Y = X\beta + \varepsilon, \tag{1}$$

où $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim \mathcal{N}_n\big(0, \tau^2 I_n\big)$, $X = (X_1, \ldots, X_n)^T \in \mathbb{R}^{n \times p}$ et $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$. L'exposant T désigne la transposition. La matrice I_n est la matrice identité de taille n et le paramètre τ caractérise le niveau de bruit. Nous notons $\Sigma = X^T X/n$ la matrice de Gram associée à la matrice de design X, $\hat{\sigma} = X^T Y/n$ et $\sigma = \mathbb{E}[\hat{\sigma}] = \Sigma\beta$.

Notre objectif est d'étudier les performances en prédiction de la régression par Moindres Carrés Partiels (Partial Least Squares, PLS). L'idée de la régression PLS est de rechercher un nombre fixe de directions, $K \in \{1, \ldots, p\}$, formées par des combinaisons linéaires des coordonnées de X, qui sont fortement corrélées avec la variable cible Y (voir Mateos-Aparicio (2011) pour une introduction complète). Le paramètre β est ensuite estimé en utilisant ces composantes.

2 La régression PLS

La régression PLS a été principalement développée dans la communauté de la chimiométrie (Martens et Naes, 1992). Cette approche a démontré sa capacité à prédire des modèles de régression avec de nombreuses variables prédictives (Garthwaite, 1994).

La méthode PLS est un algorithme itératif. L'idée est de rechercher les composantes de X les plus corrélées au vecteur Y. En particulier, pour chaque $k \in \{1, \ldots, K\}$, nous résolvons

$$\mathbf{w}_k = \underset{w \in \mathbb{R}^p}{\operatorname{argmin}} \left[-\frac{1}{n}(-Y^T X^{(k)} w) \right],$$

où $\mathbf{X}^{(k)}$ est défini itérativement comme $\mathbf{X}^{(k+1)}{=}\mathbf{X}^{(k)} - \mathbf{P}_{[\mathbf{t}_k]}(\mathbf{X}^{(k)})$ où $\mathbf{t}_k = \mathbf{X}^{(k)}\mathbf{w}_k$ et $\mathbf{P}_{[\mathbf{t}_k]}$ désigne l'opérateur de projection orthogonale sur $[\mathbf{t}_k]$. Nous renvoyons à Helland (1990) ou Castelli, Gannaz et Marteau (2024) pour une description détaillée de l'algorithme. L'estimateur PLS de β est égal à $\hat{\beta}_K = \hat{\beta}_W = W(W^T \Sigma W)^{-1} W^T \hat{\sigma}$.

Bien que le principe des Moindres Carrés Partiels (PLS) ait attiré beaucoup d'attention au fil des ans, peu de résultats théoriques ont été obtenus. Entre autres, on peut citer Helland (1990) qui a caractérisé l'espace $[W]$ résultant de l'approche PLS à l'aide de Σ et σ. Alors que Chun et Keleş (2010) ont prouvé l'inconsistance de l'estimateur PLS lorsque le nombre de covariables est trop important, Cook et Forzani (2017) ainsi que Cook et Forzani (2019) ont établi - sous des contraintes fortes sur β et la matrice de design X - le comportement asymptotique de l'erreur quadratique moyenne de prédiction et ont démontré qu'elle peut tendre vers 0 lorsque le nombre d'observations tend vers l'infini. Par ailleurs, des travaux ont également proposé de prendre en compte des contraintes de parcimonie dans l'algorithme PLS. Nous renvoyons par exemple à Durif et al. (2017) et Alsouki et al. (2023).

3 Contributions

Nous résumons ici les résultats obtenus dans Castelli, Gannaz et Marteau (2024). La forme itérative de l'algorithme PLS rend l'analyse statistique de la méthode PLS difficile.

Néanmoins, Helland (1990) a démontré que $[W] = \hat{\mathcal{G}}$, où $\hat{\mathcal{G}}$ désigne l'espace de Krylov défini comme suit :

$$\hat{\mathcal{G}} := [\hat{G}] \quad \text{avec} \quad \hat{G} = (\hat{\sigma}, \Sigma\hat{\sigma}, \ldots, \Sigma^{K-1}\hat{\sigma}).$$

Cette perspective est mieux adaptée pour évaluer les performances théoriques de $\hat{\beta}_{PLS}$. Elle fournit une formule explicite des K directions du sous-espace couvert par les poids sans référence à leur aspect itératif. En particulier, la prédiction associée à l'algorithme PLS s'écrit $X\hat{\beta}_K = P_{[XW]}Y = P_{[X\hat{G}]}Y$. Nous avons dans ce cadre une formule explicite de l'estimateur PLS :

$$\hat{\beta}_K = \operatorname{argmin}_{\beta\in\hat{\mathcal{G}}}\|Y - X\beta\|^2 = \hat{G}\hat{\Theta}^{-1}\hat{G}^T\hat{\sigma} \quad \text{où} \quad \hat{\Theta} = \hat{G}^T\Sigma\hat{G} = (\hat{\sigma}^T\Sigma^{i+j-1}\hat{\sigma})_{i,j=1..K}. \tag{2}$$

Nous introduisons également

$$G = (\sigma, \Sigma\sigma, \ldots, \Sigma^{K-1}\sigma), \quad \Theta = G^T\Sigma G, \quad D = \operatorname{diag}(\Theta), \quad R = D^{-\frac{1}{2}}\Theta D^{-\frac{1}{2}}, \quad \text{et} \quad \Lambda = \Theta^{-1}G^T\sigma. \tag{3}$$

La matrice R est la version normalisée de Θ qui peut être interprétée comme une matrice de corrélation entre les composantes de Krylov. Nous supposerons dans la suite que R est inversible.

Étant donné que l'estimateur $\hat{\beta}_{PLS}$ dans (2) implique une version estimée de Θ, il apparaît que les performances se détérioreront lorsque la plus petite valeur propre $\rho_{\min}(R)$ de R sera trop faible. Pour contourner ce problème, nous introduirons une étape de régularisation Ridge dans l'estimateur PLS. En particulier, nous examinerons l'estimateur $\hat{\beta}_{K,\alpha}$ défini comme suit

$$\hat{\beta}_{K,\alpha} = \hat{G}(\hat{\Theta} + \Delta_\alpha)^{-1}\hat{G}^T\hat{\sigma} \quad \text{pour tout} \quad \Delta_\alpha = \operatorname{diag}(\alpha_1, ..., \alpha_K) \in \mathbb{R}^K. \tag{4}$$

Nous prouvons qu'avec un choix approprié de α, l'estimateur $\hat{\beta}_{K,\alpha}$ améliore les performance de l'estimateur PLS.

Théorème 3.1. *Soit $\delta \in (0,1)$. Posons*

$$\alpha_i = c_{\delta,K}K\frac{\tau^2}{n}\rho(\Sigma)^i\operatorname{Tr}(\Sigma^i) \quad \forall i \in \{1, \ldots, K\} \quad \text{with} \quad c_{\delta,K} = 16C_\delta \ln(K+1), \tag{5}$$

où C_δ est explicite et ne ne dépend que de δ. Alors avec probabilité supérieure à $1-\delta$,

$$\frac{1}{n}\|X\hat{\beta}_{K,\alpha} - X\beta\|^2 \le \frac{2}{n}\inf_{v\in[G]}\|X(\beta - v)\|^2 + D'_{\delta,R,K}\frac{\tau^2}{n}\max\Big(\operatorname{Cond}(D)\|\Lambda\|^2 K\sum_{i=1}^K \rho(\Sigma^i)\operatorname{Tr}(\Sigma^i), \sqrt{\operatorname{Cond}(D)\|\Lambda\|^2\sum_{i=1}^K \operatorname{Tr}(\Sigma^i)^2}\Big).$$

où

$$D'_{\delta,R} = c'_\delta \operatorname{Cond}(R)^4 \ln(K+1)^2, \tag{6}$$

avec c'_δ constante positive dépendant uniquement de δ.

Le Théorème 3.1 fournit une borne non asymptotique sur la perte quadratique en prédiction, sans hypothèse sur la norme des composantes PLS autre que la dimension du sous-espace $\mathcal{G}$. Nous renvoyons à Castelli, Gannaz et Marteau (2024) pour la preuve de ce résultat. Le premier terme correspond au biais. Il mesure la distance entre le vrai signal $X\beta$ et la prédiction la plus précise dans le sous-espace de Krylov $[G]$. Ce terme dépend de K par le biais de la dimension de $\mathcal{G}$. En particulier, l'utilisation d'un grand nombre de composantes PLS (c'est-à-dire une grande valeur de K) permettra de fournir une bonne approximation (comprise comme une approximation associée à un faible biais). En revanche, l'utilisation d'un petit nombre de composantes PLS peut conduire à la situation où l'espace de Krylov $[G]$ ne peut pas fournir une bonne approximation de la cible $X\beta$. Le second terme apparaissant dans la valeur efficace de la limite mesure la variabilité de l'estimateur et peut être considéré comme un terme de variance. Il dépend essentiellement de quatre quantités : la plus petite valeur propre de R qui mesure la corrélation entre les composantes de Krylov, $\mathrm{Cond}(D)$ qui mesure la différence de normes entre les composantes de Krylov, la trace d'une puissance de la matrice de Gram Σ et la norme des composantes de Krylov Λ introduite dans (3).

Remarquons enfin que nous considérons des covariables X déterministes, tandis que Cook et Forzani (2019) considèrent des covariables X gaussiennes. Nous pensons que le fait de conserver X dans l'étude permet de mieux comprendre les mécanismes de la régression PLS, notamment par la condition de borne inférieure de la norme des composantes.

Simulations

On considère $X \in \mathbb{R}^{200\times 5}$, pour un β donné la réponse Y est générée selon (1) avec $\tau^2 = 1$. Nous définissons β par $\beta = \eta \cdot (v_1 + v_2) \in \mathbb{R}^5$ avec v_1, v_2 vecteurs propres associées à la matrice $\frac{1}{n}X^TX$ avec $n = 200$. Ainsi l'espace sous-jacent $\mathcal{G}$ est de dimension 2 et élimine la présence de biais. Le paramètre η peut être vu comme une mesure de rapport signal sur bruit. Nous générons $N = 2000$ échantillons associées à chaque valeur de η. Les covariables X sont fixes pour les N échantillons tandis que le bruit ε varie. Nous appliquons la PLS avec $K = 2$ composantes en considérant l'estimateur PLS $\hat{\beta}_K$ et sa version régularisée $\hat{\beta}_{K,\alpha}$ définis respectivement en (2) et (4). Pour souligner l'effet de l'inversion de $\hat{\Theta}$ sur le processus d'estimation, nous incluons également dans l'analyse le pseudo-estimateur $\hat{\beta}^{or}$ défini comme suit :

$$\hat{\beta}^{or} = \hat{G}\Theta^{-1}G^T\sigma. \tag{7}$$

Le pseudo-estimateur $\hat{\beta}^{or}$ correspond au cas spécifique où les coordonnées de Krylov Λ sont supposées connues et peut être considéré comme un oracle. Il est linéaire dans la direction du sous-espace $\hat{\mathcal{G}}$. Le risque quadratique associé à $\hat{\beta}^{or}$ ne dépend pas du paramètre η. Enfin, nous étudions les performances des estimateurs par l'évaluation de

$$\mathrm{MSE}_{\eta,j} = \frac{1}{n \times N}\sum_{i=1}^{N} \|X(\beta_\eta - \hat{\beta}_{i,j})\|^2,$$

où $\hat{\beta}_{i,j}$ est l'estimateur selon le i-ème de $Y = X\beta + \varepsilon$ avec j indiquant le choix de l'estimateur. L'indice $j = 1, 2, 3$ correspond respectivement à l'estimateur PLS (2), à l'estimateur oracle

(7) et à l'estimateur Ridge (4).

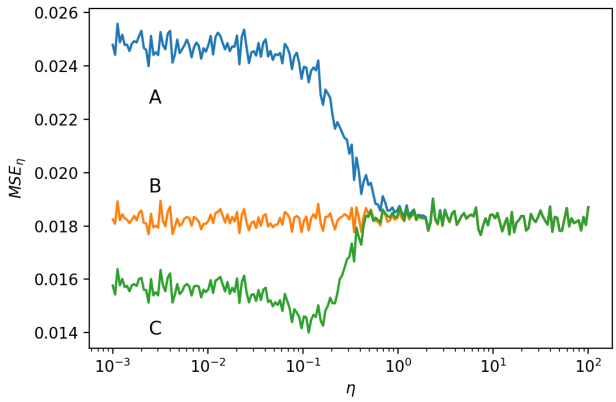


Figure 1: MSE l'estimateur PLS (A), l'estimateur Oracle (B) et l'estimateur PLS régularisée (C), en fonction du paramètre η.

La figure 1 montre que, lorsque le paramètre du rapport signal/bruit η est faible, la qualité de l'estimateur PLS se détériore. Dans ces conditions, les avantages de la régularisation Ridge sont notables. En particulier, lorsque Θ est très mal conditionnée (pour les petits η).

L'estimateur oracle (7) correspond à un estimateur dans lequel l'axe PLS Λ est connu et seules les coordonnées de β sur l'axe sont estimées. Comme l'erreur de prédiction est constante, cela montre que la qualité de l'estimation dépend principalement de la qualité de l'axe estimé. En particulier, la dégradation de la PLS pour les faibles η est basée sur l'estimation de Λ, principalement à travers l'erreur sur $\hat{\Theta}^{-1}$. La régularisation Ridge améliore cette estimation.

Conclusion

Nos résultats établissent des bornes non asymptotiques sur la perte quadratique en prédiction de l'estimateur PLS régularisée. Une régularisation Ridge est proposée pour surmonter les scénarios où la matrice Θ peut-être très mal conditionnée. La prise en compte d'un cadre non asymptotique, et de covariables non aléatoires, permet de mettre en évidence que la procédure est efficace. Basée sur la représentation de Krylov de l'estimateur PLS, cette approche requiert seulement que les composantes de Krylov sont linéairement indépendantes. De plus, nos travaux mettent en évidence l'influence de la matrice de Gram des covariables Σ. Nous adoptons le point de vue de l'espace de Krylov qui est un cadre approprié pour étudier les performances théoriques. Cette approche nous permet d'appliquer des résultats de déviation et de fixer des bornes de prédiction. Cette méthode nous permet de garantir les performances de la régression PLS régularisée dont le paramètre α ne dépend que de la dimension, du bruit et de la matrice de Gram Σ.

Enfin, une étude par simulation montre que l'hypothèse d'un rapport signal/bruit suffisant

pour assurer la qualité de l'approche PLS est judicieuse. Elle montre également que la régularisation de Ridge permet de surmonter cette hypothèse.

Bibliographie

Alsouki L., Duval L., Marteau C., El Haddad R. and Wahl F. (2023) Dual-sPLS: a family of Dual Sparse Partial Least Squares regressions for feature selection and prediction with tunable sparsity; evaluation on simulated and near-infrared (NIR) spectra, *Chemometrics and Intelligent Laboratory system*, 237, 104813.

Bickel, Peter J. and Ritov, Ya'acov and Tsybakov, Alexandre B. (2009) Simultaneous analysis of lasso and Dantzig selector *The Annals of Statistics*, 37, 1705–1732.

Castelli L., Gannaz I. and Marteau M. (2023) A non asymptotic analysis of the single component PLS regression, arXiv preprint arXiv:2310.10115.

Castelli L., Gannaz I. and Marteau M. (2024) A non asymptotic upper bound in prediction for the PLS estimator, arXiv preprint arXiv:2410.10237.

Chun H. and Keles S. (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1), 3-25.

Cook, R. D. and Forzani L. (2017) Big data and partial least-squares prediction, *Canadian Journal of Statistics*, 46(1), 62-78.

Cook, R.D. and Forzani L. (2019) Partial least squares prediction in high-dimensional regression, *The Annals of Statistics*, 47(2), 884-908..

Durif G., Modolo L., Michaelsson J., Mold J.E., Lambert-Lacroix S. and Picard F. (2017) High dimensional classification with combined adaptive sparse PLS and logistic regression, *Bioinformatics*, 34(3), 485-493.

Garthwaite P.H. (1994) An interpretation of partial least squares, *Journal of the American Statistical Association*, 89(425), 122-127.

Helland I. (1990) Partial least squares regression and statistical models, *Scandinavian journal of statistics*, 17(2), 97-114.

Martens H. and Naes T. (1992) Multivariate calibration, *John Wiley & Sons*.

Mateos-Aparicio G. (2011) Partial least squares (PLS) methods: origins, evolution, and application to social sciences, *Communications in Statistics - Theory and Methods*, 40(13), 2305-2317.

Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

Variational autoencoder approach for inference of nonlinear mixed effect models based on ordinary differential equations

Zhe Li [1] & Rodolphe Thiébaut [1] & Mélanie Prague [1] & Quentin Clairon [1]

University of Bordeaux, Department of Public Health, Inserm Bordeaux Population Health Research Centre, Inria SISTM team, VRI, France

zhe.li@u-bordeaux.fr, quentin.clairon@u-bordeaux.fr

Résumé. Nous proposons une méthode basée sur les réseaux de neurones pour estimer les paramètres des modèles non linéaires à effet mixtes basés sur les équations différentielles ordinaires (NLME-ODE) à partir de données longitudinales. Pour ce problème d'inférence, les approches classiques par maximisation de la vraisemblance s'appuyant sur l'algorithme SAEM, basé sur une procédure de Monte-Carlo Markov Chain (MCMC) pour reconstruire les distributions a posteriori des paramètres spécifiques aux sujets, offrent des estimations précises pour des problèmes de dimensions raisonnables. Cependant, lorsque la dimension du modèle ou des données augmente et que les observations disponibles par sujet deviennent rares, de telles procédures d'inférence peuvent se dégrader en raison de problèmes de convergence des approches MCMC. Pour pallier ces limites, nous adoptons une approche variationnelle bayésienne, où nous remplaçons la distribution a posteriori des paramètres spécifiques aux sujets par une approximation optimisée via l'ELBO (Evidence Lower Bound). Dans un cadre d'apprentissage profond, cette approximation est réalisée à l'aide de réseaux neuronaux, conduisant à un modèle d'auto-encodeur variationnel. Cette reformulation transforme l'inférence par échantillonnage en un problème d'optimisation, résoluble par descente de gradient, permettant ainsi de traiter des modèles complexes sans les coûts computationnels et l'instabilités potentielles des MCMC. Enfin, nous estimons la variance des paramètres via une approximation de l'information de Fisher observée et vérifions l'absence de nouveaux problèmes d'identifiabilité causés par l'introduction de paramètre de nuisance additionnels, garantissant ainsi la robustesse et l'interprétabilité des résultats.

Mots-clés. Modèles non-linéaire à effet mixte, inférence variationnelle, réseaux de neurone, apprentissage profond

Abstract. We propose a neural network-based method for estimating parameters in nonlinear mixed effects models based on ordinary differential equations (NLME-ODE), using population data collected from multiple subjects. For this inference problem relying on likelihood maximization, the Stochastic Approximation of Expectation Maximization (SAEM) algorithm, which uses a Monte-Carlo Markov Chain (MCMC) procedure to reconstruct the posterior distributions of subject-specific parameters, provides accurate estimates for the problems of reasonable dimensions. However, when model/data dimension increases or available data per subject becomes sparse, such inference procedures can face degradation due to MCMC convergence issues. Therefore, alternative methods need to be proposed to regularize the inverse problem of parameter estimation in such complex settings, which

are becoming increasingly relevant for real data analysis, such as in clinical trials. We use the Bayesian variational approach to construct an approximation (a lower bound) of the data likelihood, leading to the Evidence Lower Bound (ELBO) of the marginal likelihood, which serves as the new objective function. Such approaches involve replacing the true subject-specific parameters (which act as latent variables) with an approximation, chosen from a family of candidate distributions, to estimate their posterior distribution. In our deep-learning framework, this family is constituted of parametrized neural networks, thus leading to the so-called variational auto-encoder setting. The idea is to transform sampling-based inference procedure, which can be intractable, into an optimization problem that can be solved with the gradient-based methods, among others. In this way, we bypass the sampling step in the MCMC based approaches in order to scale up into high dimensional/complex settings without suffering from too much computational cost and inaccurate estimation regarding the reconstructed a posterior law of the subject specific parameters. Once the point-wise estimation is done, we proceed to the variance estimation via Observed Fisher Information approximation. Moreover, we check that our approximated method (which requires the introduction of additional nuisance parameters) does not come with the prize of the identifiability issues which are highly critical for this inverse problem relying on parameter estimation interpretability.

Keywords. Nonlinear mixed-effect models, variational inference, neural network, mechanistic modeling

1 Introduction

1.1 NLME-ODE Models

We focus on a population of n subjects in which the available data for the i-th subject are given by the observation model:

$$Y_{ij} = h_{\theta_i}(X_{\theta_i}(t_{ij})) + \epsilon_{ij} \quad \text{with} \quad \epsilon_{ij} \sim \mathcal{N}(0, \Sigma_\epsilon) \quad \text{and} \quad i = 1, 2, ...n; \quad j = 1, 2, ..., n_i$$

at the n_i timepoints $t_i = (t_{i_1}, ..., t_{i_{n_i}})$ linking the raw data to the variable X_{θ_i} assumed to be the solution of the ODE:

$$\begin{cases} \dot{X_{\theta_i}} = f_{\theta_i}(X_{\theta_i}, t) \\ X_{\theta_i}(0) = x_{i,0}(\theta_i) \end{cases}$$

requiring the introduction of structural parameter θ_i for its definition.

To address both inter-subject variability and maintain homogeneity at the population level, especially when dealing with sparse data per subject, we adopt a mixed-effects framework:

$$\theta_i = \theta e^{b_i} \text{ with } b_i \sim \mathcal{P}_\phi(b_i) = \mathcal{N}(0, \Omega_b^2) \tag{1}$$

θ represents the fixed effect, common to all subjects, while b_i captures subject-specific deviations, assumed to follow a normal distribution. So, we are left with the parameter estimation problem of $\phi = (\theta, \Sigma_\epsilon, \Omega_b)$.

1.2 Classical method

One popular method for ϕ estimation is via the maximization of the likelihood (or equivalently log-likelihood) of the observation, given by:

$$LL(\phi) = \sum_{i=1}^{n} \ln \mathbb{E}_{b_i \sim \mathcal{P}_\phi(b_i)} \left[\mathcal{P}_\phi(Y_i|b_i) \right] \tag{2}$$

with

$$\mathbb{E}_{b_i \sim \mathcal{P}_\phi(b_i)} \left[\mathcal{P}_\phi(Y_i|b_i) \right] = \int \mathcal{P}_\phi(Y_i|b_i) \mathcal{P}_\phi(b_i) \, db_i.$$

The optimization of $LL(\phi)$ based on its direct computation is generally intractable due to the high-dimensional integral involved in the previous expectation definition. This calls for more indirect/approximated approaches. The currently most widespread one being the SAEM algorithm (proposed by Delyon and Lavielle (1999)) replaces the original $LL(\phi)$ optimization problem by the construction of a sequence of functions Q_l and updated sequence of estimators ϕ_l recursively defined by:

$$\begin{cases} Q_l(\phi|\phi_l) = Q_{l-1}(\phi|\phi_{l-1}) + \gamma_l \left(\frac{1}{\tilde{M}} \sum_{m=1}^{\tilde{M}} \log \mathcal{P}_\phi(Y_i, b_i^{(m)}) - Q_{l-1}(\phi|\phi_{l-1}) \right) \\ \phi_l = \arg\max Q_l(\phi \mid \phi_l) \end{cases} \tag{3}$$

where the $b_i^{(m)}$ is sampled from $\mathcal{P}_{\phi_{l-1}}(b_i|Y_i)$, the posterior distribution of the subject specific parameter which is generally unknown. That is why Kuhn and Lavielle (2004) proposed relying on the Markov chain Monte Carlo (MCMC) procedure to approximate it.

However, in complex statistical models, such as those involving sparse data, high-dimensional latent variables, and potential model misspecification, MCMC methods, although widely used, can face significant challenges that may hinder their performance or even cause failure. In sparse settings, there is limited information to guide the inference process, making it difficult for MCMC to efficiently explore the posterior. As the number of latent variables increases, the posterior distribution becomes more complex and high-dimensional. MCMC methods often struggle in high-dimensional spaces because they require many steps to explore the full distribution. When the model is misspecified, like incorrect priors, the posterior distribution may fail to reflect the true underlying data distribution. This can lead to biased or inconsistent parameter estimates, and MCMC may fail to converge to the correct posterior.

2 Method

2.1 Auto-encoder Bayesian variational approach

An alternative procedure would be to use an approximation $\mathcal{P}_\phi(b_i \mid Y_i) \approx \mathcal{Q}_\psi(b_i \mid Y_i)$ where the proxy distribution $\mathcal{Q}_\psi(b_i \mid Y_i)$ is chosen among a family of candidate distributions. In this way,the sampling-based inference is transformed into optimization to enable scalable, efficient estimation in high-dimensional settings. We can control the level of regularity of

$b_i \sim \mathcal{Q}_\psi(b_i|Y_i)$ and account for model errors by relaxing the constraint imposed by the original model. Similarly as in the SAEM approach, we also construct an alternative cost function hoped to be more suited for optimization purpose than the initial likelihood. In this case, it is given by the so called ELBO :

$$\text{ELBO} = \mathbb{E}_{\mathcal{Q}_\psi(b_i|Y_i)}[\log \mathcal{P}_\phi(Y_i \mid b_i)] - D_{KL}(\mathcal{Q}_\psi(b_i|Y_i)||\mathcal{P}_\phi(b_i))$$

As often done for regularization purpose, the ELBO cost function can be seen as a trade-off between a data fidelity term ($\mathbb{E}_{\mathcal{Q}_\psi(b_i|Y_i)}[\log \mathcal{P}_\phi(Y_i \mid b_i)]$) which prioritizes capturing patterns in the data accurately and a penalty term ($D_{KL}(\mathcal{Q}_\psi(b_i|Y_i)||\mathcal{P}_\phi(b_i))$) which encourages the latent variables to remain consistent with the prior distribution $\mathcal{P}_\phi(b_i)$ which is often assumed as a Gaussian distribution as shown in equation (1). We are left with the choice of acceptable set of candidate distribution for $\mathcal{Q}_\psi$, for this we choose the Gaussian distribution family $\mathcal{Q}_\psi(b_i \mid Y_i) = \mathcal{N}(\Gamma_\psi(Y_i), \Sigma_\psi(Y_i))$ since it allows to derive a closed form for the Kullback Leibler distance thus enhancing computational speed. The mean function Γ_ψ and variance function Σ_ψ are parameterized by the neural network giving rise to variational auto-encoder (Kingma and Welling [2013]). Our encoder is built on one block of Convolution, Batch normalization and Rectified Linear Unit ($ReLU$) layers.

2.2 Identifiability

Our inference problem belongs to the class of learning generative model problem. In generative modeling, identifiability is often overlooked, but it is a critical concern in our modeling situation where parameter interpretability is crucial for statistical inference. Our approximation method with the neural network incorporates extra nuisance parameters into the model through the encoder. To ensure identifiability is not compromised, we execute the parameters estimation several times, with different, randomly generated, initial values of fixed effects. The goal is to assess the robustness of the convergence.

2.3 Estimation of the variance of the parameters

Another issues rarely considered in generative model is the quantification of uncertainty of the obtained estimation which is still crucial in our statistical setting. For this task, we evaluate the variance of the estimated parameters as well as the coverage rate of the derived 95% confidence interval. We use the inverse of the observed information matrix as asymptotic variance covariance estimator $I_n(\hat{\phi})$:

$$I_n(\hat{\phi}) = -\nabla^2_\phi \ln \mathcal{P}_{\hat{\phi}}(Y_i) = -\frac{\partial}{\partial \phi} \nabla_\phi \ln \mathcal{P}_{\hat{\phi}}(Y_i) = -\frac{\nabla^2_\phi \mathcal{P}_{\hat{\phi}}(Y_i)\mathcal{P}_{\hat{\phi}}(Y_i) - \nabla_\phi \mathcal{P}_{\hat{\phi}}(Y_i) \nabla_\phi \mathcal{P}_{\hat{\phi}}(Y_i)^T}{\mathcal{P}_{\hat{\phi}}(Y_i)^2} \tag{4}$$

The quantities $\mathcal{P}_{\hat{\phi}}(Y_i)$, $\nabla_\phi \mathcal{P}_{\hat{\phi}}(Y_i)$ and $\nabla^2_\phi \mathcal{P}_{\hat{\phi}}(Y_i)$ appearing in equation (4) are approximated via Monte-Carlo procedure and reparametrization trick which is an alternative method for generating samples. It is often possible to express the random variable b as a deterministic variable $b = g_\phi(\epsilon)$, where ϵ_l is an auxiliary variable with independent marginal $\mathcal{P}(\epsilon_l)$, and $g_\phi(\cdot)$ is some vector-valued function parameterized by ϕ (Kingma and Welling [2013]).

3 Illustration

To illustrate the relevancy of inference strategy, we consider a partially observed two-compartments pharmacokinetic model with first-order absorption and elimination.

$$\begin{cases} \dot{X_{1,i}} = \phi_{2,i}X_{2,i} - \phi_{1,i}X_{1,i} \\ \dot{X_{2,i}} = -\phi_{2,i}X_{2,i} \\ (X_{1,i}(0), X_{2,i}(0)) = (x_{1,0}, x_{2,0}) \end{cases} \quad \text{with} \quad \begin{cases} \log(\phi_{1,i}) = \theta_1 + b_i \\ \log(\phi_{2,i}) = \theta_2 \end{cases} \quad b_i \sim \mathcal{N}(0, \omega_b^2)$$

We generated 100 simulation replicates with 100 participants corrupted with measurement error. The true population parameter values are $\theta^* = (\theta_1^* \ , \ \theta_2^*) = (log(0.5) \ , \ log(2))$ and the standard deviation of the random effect b_i is $\sigma_b^\star = 0.5$. We are in a partially observed framework where only $X_{1,i}$ is accessible. The initial conditions are $(x_{1,0}, x_{2,0}) = (2, 3)$. We generated $n_i = 11$ observations per subject on $[0, T] = [0, 10]$ with Gaussian measurement noise of standard deviation $\sigma_\epsilon = 0.2$. From these replicates, we estimate the parameters $\phi = (\theta_1, \theta_2, \sigma_b)$ and then we calculate the mean, the relative bias, the empirical variance as well as the relative root mean square error of our estimator to quantify its practical accuracy. We also estimate the variance derived from the Observed Fisher Information Matrix as well as the coverage rate of the corresponding 95% confidence intervals. The estimation results are given in table 1. We can see that both empirical and estimated variance globally coincides and the actual coverage rate is generally close to the expected rate of 0.95, thus indicating a well-conditioned estimation problem. We compared our results with the estimations by the algorithm SAEM given in table 2. We can see that, in terms of accuracy, our approach gives similar performance comparing to SAEM algorithm's estimation. We also ensured that our approach does not lead to new identifiability issues, as shown in Figure 1.

Parameter	**Relative Bias**	**Empirical Variance**	**Estimated Variance**	**Coverage Rate**
θ_1	0.0050	0.0023	0.0024	0.92
θ_2	-0.0033	0.0016	0.0015	0.94
$\log\left(\omega_b^2\right)$	-0.0230	0.0191	0.0237	0.92

Table 1: **ELBO** point estimation quality and variance estimation

Parameter	**Relative Bias**	**Empirical Variance**	**Estimated Variance**	**Coverage Rate**
θ_1	0.0130	0.0024	0.0025	0.94
θ_2	-0.0031	0.0016	0.0032	0.92
$\log\left(\omega_b^2\right)$	-0.0099	0.0204	0.0208	0.95

Table 2: **SAEM** point estimation quality and variance estimation

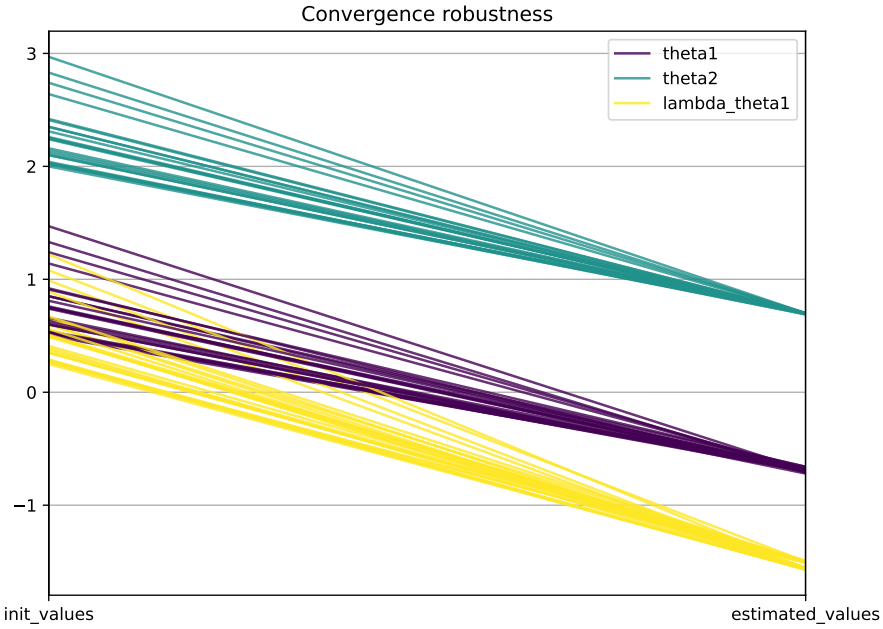


Figure 1: PKPD convergence robustness

4 Conclusion

The performance of the resulting algorithm is encouraging to move to a more complex configuration, with stronger nonlinearity and modeling errors.

Bibliographie

Quentin Clairon, Mélanie Prague, Delphine Planas, Timothée Bruel, Laurent Hocqueloux, Thierry Prazuck, Olivier Schwartz, Rodolphe Thiébaut, Jérémie Guedj. Modeling the kinetics of the neutralizing antibody response against SARS-CoV-2 variants after several administrations of Bnt162b2. PLoS Computational Biology 19 (8), e1011282

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. Proceedings of the 2nd International Conference on Learning Representations.

B. Delyon, M. Lavielle and E. Moulines, Convergence of a stochastic approximation version of the EM algorithm. Ann. Statist. 27 (1999) 94–128.

Estelle Kuhn and Marc Lavielle. Coupling a stochastic approximation version of em with an mcmc procedure. ESAIM: Probability and Statistics, 8:115–131, 2004

Lixoft. Monolix version 2024r1. Antony, France, 2024.

Régression géodésique intrinsèque dans des groupes de Lie : modèle et propriétés statistiques

Johan Aubray[1], Florence Nicol[1], Stéphane Puechmorel[1] & Anne-Françoise Yao[2]

[1] *Université de Toulouse, Laboratoire ENAC, France, prenom.nom@enac.fr*
[2] *Université Clermont Auvergne, Laboratoire de Mathématiques Blaise Pascal et CMAP, Ecole polytechnique, Palaiseau, France, anne.yao@uca.fr*

Résumé. Dans cette communication, nous nous intéressons au problème de la régression géodésique dans un groupe de Lie dans le but d'estimer la position d'un mobile à partir de positions mesurées bruitées. Dans le contexte de la gestion du trafic aérien (ATM), l'estimation de la position d'aéronefs ou d'un drone peut s'avérer complexe, les données ne pouvant être traitées avec des outils statistiques usuels comme c'est le cas de données appartenant à un espace euclidien. Dans Aubray et Nicol (2024), ce problème de régression a été traité en modélisant l'évolution d'un mobile (position et orientation) dans le groupe Spécial Euclidien $SE(n)$ avec $n = 3$. La géodésique s'ajustant le mieux aux mesures de la trajectoire d'un mobile au sens des moindres carrés riemanniens a été calculée en utilisant la connexion de Levi-Civita. Dans cette contribution, nous étudions les propriétés statistiques des estimateurs intrinsèques de la régression géodésique.

Mots-clés. Statistique sur des variétés, régression géodésique, variétés riemanniennes, groupes de Lie, $SE(3)$, gestion du trafic aérien, connexion, géométrie différentielle

Abstract. In this paper, we focus on the problem of geodesic regression in a Lie group, with the aim of estimating the position of a mobile from noise-measured positions. In the context of air traffic management (ATM), estimating the position of aircraft or a drone can be a complex task, as the data cannot be processed with the usual statistical tools as in the case of data belonging to a Euclidean space. In Aubray et Nicol (2024), this regression problem has been tackled by modeling the evolution of a mobile (position and orientation) in the Special Euclidean group $SE(n)$ with $n = 3$. The geodesic that best fits the measurements of a mobile's trajectory in the sense of Riemannian least squares was calculated using the Levi-Civita connection. In this contribution, we study the statistical properties of intrinsic estimators of geodesic regression.

Keywords. Statistics on manifolds, geodesic regression, Riemannian manifolds, Lie groups, $SE(3)$, air traffic management, connection, differential geometry

1 Introduction et contexte

Nous nous intéressons au problème de l'estimation de la position d'un aéronef à partir de données bruitées. Dans le contexte de la gestion du trafic aérien (ATM), l'estimation de

la position d'aéronefs ou d'un drone peut s'avérer complexe, les données ne pouvant être traitées avec des outils statistiques usuels. La géométrie riemannienne fournit un cadre de travail pertinent pour étudier de telles données : celles-ci sont vues comme des points appartenant à une variété différentielle, permettant ainsi d'écrire le problème dans une autre espace de représentation. Dans notre contexte, plutôt que de considérer les trajectoires dans un espace d'état, la position et l'orientation d'un mobile peuvent être représentées dans un autre espace de représentation, en particulier dans celui des groupes de Lie.

Partant du problème de régression géodésique développé dans des variétés riemanniennes par Fletcher (2011), nous avons examiné ce problème dans Aubray et collab. (2023) en modélisant l'évolution d'un mobile (position et orientation) dans l'espace des groupes de Lie, plus particulièrement le groupe Spécial Euclidien $SE(3)$. Les calculs de géodésiques se faisaient alors, conformément à l'article original, avec la connexion de Levi-Civita, qui a pour intérêt principal sa compatiblité avec la métrique de la variété. Dans Aubray et collab. (2024) et Aubray et Nicol (2024), ce principe a été étendu à une connexion quelconque. En effet, dans bon nombre d'applications, on utilise d'autres connexions que celle de Levi-Civita. Par exemple, Pennec (2013) utilise les connexions de Cartan-Schouten pour exprimer les géodésiques dans les groupes de Lie comme sous-groupes à un paramètre. De même, les α-connexions (Amari et Nagaoka (2000)) sont souvent utilisées dans des modèles statistiques sur des variétés. L'intérêt d'utiliser une connexion quelconque peut se manifester de deux manières : soit elle permet de simplifier l'écriture du problème, soit elle permet d'exprimer une quantité non intrinsèque, c'est-à-dire dont l'écriture dépend d'un système de coordonnées spécifique (e.g. c'est le cas des courbes polynomiales qui s'expriment différement selon le système de coordonnées choisi).

Dans la contribution présente, nous examinons le problème d'un point de vue statistique et étudions les propriétés du minimiseur de l'erreur quadratique moyenne et de sa version empirique.

2 Le problème de la régression géodésique

2.1 Élements de géométrie différentielle

Cette section est destinée à préciser quelques notions de géométrie différentielle qui seront utilisés plus loin. On pourra consulter do Carmo et Flaherty (2013), Gallier (2011) et Gallot et collab. (2004) pour plus de détails.

Soit $\mathcal{M}$ une variété différentielle de dimension finie, $T\mathcal{M}$ son fibré tangent et $E \xrightarrow{\pi} \mathcal{M}$ un fibré vectoriel de base $\mathcal{M}$. On notera $\Gamma(E)$ le $C^\infty(\mathcal{M})$-module des sections lisses de ce fibré. Une connexion affine ∇ est un $\mathbb{R}$-opérateur linéaire:

$$\nabla : \Gamma(E) \to \Gamma(T^\star\mathcal{M} \otimes E) \tag{1}$$

vérifiant l'identité de Liebniz:

$$\forall f \in C^\infty(\mathcal{M}),\ \forall s \in \Gamma(E),\ \forall X \in T\mathcal{M},\ \nabla_X(fs) = X(f)s + f\nabla_X s \tag{2}$$

Si ∇ est une connexion affine sur le fibré tangent $T\mathcal{M}$, on dira qu'une courbe lisse $\gamma:[0,1] \to \mathcal{M}$ est une courbe géodésique de ∇ si pour tout $t \in]0,1[$, $\nabla_{\dot\gamma}\dot\gamma = 0$. Cette relation est une équation différentielle linéaire du second ordre, on en déduit que si p est un point de $\mathcal{M}$ et v un vecteur de $T_p\mathcal{M}$, il existe un intervalle $[0, t_v[$ et une unique courbe lisse $\gamma:[0,t_v[\to \mathcal{M}$ solution de $\nabla_{\dot\gamma}\dot\gamma = 0$ avec les conditions initiales $\gamma(0) = p$, $\dot\gamma(0) = v$. Finalement, par compacité de la boule unité de $T_p\mathcal{M}$, on en déduit l'existence d'un $\eta > 0$ tel que si $v \in T_p\mathcal{M}$ vérifie $|v| < \eta$, alors la géodésique γ telle que $\gamma(0) = p$, $\dot\gamma(0) = v$ est définie en $t = 1$.

Définition 2.1. *Avec les notations précédentes, l'application exponentielle* $\mathrm{Exp}_p : T_p\mathcal{M} \to \mathcal{M}$ *est définie, pour* $v \in T_p\mathcal{M}$ *tel que* $|v| < \eta$, *par:*

$$\mathrm{Exp}_p v = \gamma(1) \tag{3}$$

avec γ *la géodésique de* ∇ *vérifiant* $\gamma(0) = p$, $\dot\gamma(0) = v$.

Si $e_1, \ldots, e_n$ est une base de $T_p\mathcal{M}$, alors il existe un voisinage ouvert U de l'origine de $\mathbb{R}^n$ telle que si $(t_1, \ldots, t_n) \in U$, l'application $\mathrm{Exp}_p(t_1e_1 + \cdots + t_ne_n)$ est bijective. L'association $(t_1, \ldots, t_n) \mapsto \mathrm{Exp}_p(t_1e_1 + \cdots + t_ne_n)$ définit un système de coordonnées locales, usuellement appelées coordonnées normales.

En géométrie riemannienne, la connexion permet de définir le transport parallèle. Si $\mathcal{M}$ est une variété équipée d'une connection ∇, et $c:[0,1] \to \mathcal{M}$, on dit que le champ de vecteurs V est parallèle le long de c si pour tout $t \in [0,1]$, $\nabla_{\dot c} V(t) = 0$. On peut alors montrer le théorème suivant :

Théorème 2.1. *Soit une courbe* $c : [0,1] \to \mathcal{M}$ *et* $v \in T_{c(0)}\mathcal{M}$ *un vecteur, alors il existe un unique champ de vecteurs parallèle* $V(t)$ *le long de c, et on note* $V(t) = \Pi_0^t v$ *l'image de v par transport parallèle.*

La métrique permet de calculer des angles, des longueurs et des distances. Une courbe de longeur minimale joignant deux points p et q, est appelée géodésique minimisante; sa longueur est la distance géodésique entre p et q.

Soit $(\mathcal{M}, g)$ une variété riemanienne. Une connexion ∇ est dite métrique ssi pour tous champs de vecteurs (X, Y, Z) sur $\mathcal{M}$, on a :

$$(\nabla_Z g)(X,Y) = \nabla_Z(g(X,Y)) - g(\nabla_Z X, Y) - g(X, \nabla_Z Y) = 0. \tag{4}$$

La connexion de Levi-Civita est la seule connexion affine qui est à la fois sans torsion et compatible avec la métrique g. La métrique est alors préservée par transport parallèle. Notamment, le produit scalaire de deux vecteurs transportés parallèlement le long d'une courbe par la connexion de Levi-Civita est constant.

Dans ce travail, les variétés que nous utilisons sont des groupes de Lie. Un groupe de Lie G est une variété lisse qui est également un groupe tel que les opérations de produit et d'inversion du groupe sont lisses. Cette structure de groupe permet de se ramener en l'identité e à partir de tout point $g \in G$ par simple translation à gauche ou à droite. L'espace tangent en e, noté $\mathfrak{g}$, est appelé algèbre de Lie.

2.2 Régression géodésique dans un groupe de Lie

Dans Aubray et collab. (2023), nous avions montré que l'utilisation d'une métrique invariante à gauche permettait de réduire la problème de minimisation de la fonctionnelle

$$L(\gamma) = \frac{1}{2}\int_0^1 \langle\!\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle\!\rangle_g dt,$$

à la minimisation de la fonctionnelle réduite dans $\mathfrak{g}$:

$$L(\xi) = \frac{1}{2}\int_0^1 \left\langle \dot{\xi}(t), \dot{\xi}(t) \right\rangle dt,$$

Si G est muni de la connexion de Levi-Civita ∇^{LC}, alors une géodésique minimisante γ doit satisfaire :

$$\nabla^{LC}_{\dot{\gamma}(t)} \dot{\gamma}(t) = 0, \tag{5}$$

et l'équivalent riemannien d'un polynôme de degré p est une courbe γ telle que :

$$\left(\nabla^{LC}_{\dot{\gamma}(t)}\right)^p \dot{\gamma}(t) = 0. \tag{6}$$

Nous proposons une réécriture du modèle de régression géodésique dans une variété riemannienne exprimé dans Fletcher (2011) comme suit. Soit Y une variable aléatoire à valeurs dans $\mathcal{M}$ et T une variable aléatoire à valeurs dans $\mathbb{R}$. Notre fonction de régression, γ, est telle que $\forall t \in \mathbb{R}\ \gamma(t) = E(Y|T = t)$ où

$$\gamma(t) = \mathrm{Exp}_p(tv)$$

est la géodésique donnée par les conditions initiales $\gamma(0) = p$ et $\dot{\gamma}(0) = v$. Le modèle de régression est tel que, pour toute observation (y, t) de (Y, T),

$$y = \mathrm{Exp}_{\gamma(t)}(\epsilon), \tag{7}$$

où ϵ est une observation d'une variable aléatoire gaussienne, ε, à valeurs dans $T_{\gamma(t)}\mathcal{M}$. On cherche alors la géodésique γ qui minimise l'erreur quadratique moyenne :

$$E(\gamma) = \mathbb{E}\left(d\left(\gamma(T), Y\right)^2\right) \tag{8}$$

où d est une distance géodésique définie pour tout $\gamma(t)$ et y dans $\mathcal{M}$.

Dans Hinkle et collab. (2012), ce modèle de régression géodésique est généralisé à un modèle polynomial. L'estimation de γ se déduit de la minimisation de la version empirique de $E(\gamma)$ sous les contraintes (5) et (6). Dans Aubray et collab. (2023), ce problème de régression a été traité en modélisant l'évolution d'un mobile (position et orientation) dans le groupe Spécial Euclidien $SE(n)$ avec $n = 3$. La géodésique s'ajustant le mieux aux mesures de la trajectoire d'un mobile au sens des moindres carrés riemanniens a été calculée en utilisant la connexion de Levi-Civita. Une généralisation de ce problème en utilisant une connexion quelconque et le formalisme des formes différentielles a été proposé dans Aubray et collab. (2024) et Aubray et Nicol (2024).

Contrairement au cas Euclidien, il n'existe pas d'expressions analytiques des estimateurs des paramètres p et v de la régression géodésique. Cependant, Karcher (1977) (voir aussi Bhattacharya et Patrangenaru (2003) ou Le (2001)) a montré que $E(\gamma)$, admet un minimun global appelé moyenne intrinsèque ou moyenne de Fréchet sous l'hypothèse que $E(\gamma) < \infty$. L'objectif est de ce travail est de donner des propriétés du minimiseur de $E(\gamma)$ et de sa version empirique.

References

Amari, S.-I. et H. Nagaoka. 2000, *Methods of information geometry*, Translations of mathematical monographs 191, American Mathematical Society and Oxford University Press, ISBN 9780821843024.

Aubray, J. et F. Nicol. 2024, «Polynomial Regression on Lie Groups and Application to SE(3)», *Entropy*, vol. 26, 10, doi:10.3390/e26100825, p. 825. URL `https://hal.science/hal-04747804`.

Aubray, J., F. Nicol et S. Puechmorel. 2023, «Régression géodésique dans SE(3) : application à l'estimation de la position d'un mobile», dans *54es Journées de Statistique*, Société Française de Statistique (Sfds), Bruxelles (Belgique), Belgium. URL `https://hal.science/hal-04351972`.

Aubray, J., F. Nicol et S. Puechmorel. 2024, «Regression on a Lie group with any connection», dans *55èmes Journées de Statistiques*, Société Française de Statistique (Sfds) and Université de Bordeaux, Bordeaux (France), France. URL `https://hal.science/hal-04606135`.

Bhattacharya, R. et V. Patrangenaru. 2003, «Large sample theory of intrinsic and extrinsic sample means on manifolds», *The Annals of Statistics*, vol. 31, 1, p. 1–29.

do Carmo, M. et F. Flaherty. 2013, *Riemannian Geometry*, Mathematics: Theory & Applications, Birkhäuser Boston, ISBN 9780817634902.

Fletcher, T. 2011, «Geodesic Regression on Riemannian Manifolds», dans *Proceedings of the Third International Workshop on Mathematical Foundations of Computational Anatomy - Geometrical and Statistical Methods for Modelling Biological Shape Variability*, édité par Pennec, Xavier, Joshi, Sarang, Nielsen et Mads, Toronto, Canada, p. 75–86. URL `https://hal.inria.fr/inria-00623920`.

Gallier, J. 2011, *Basics of Manifolds and Classical Lie Groups: The Exponential Map, Lie Groups, and Lie Algebras*, Springer New York, New York, NY, ISBN 978-1-4419-9961-0, p. 459–528, doi:10.1007/978-1-4419-9961-0.

Gallot, S., D. Hulin et J. Lafontaine. 2004, *Riemannian Geometry*, Universitext, Springer Berlin, Heidelberg, ISBN 978-3-540-20493-0.

Hinkle, J., P. Muralidharan, P. T. Fletcher et S. Joshi. 2012, «Polynomial regression on riemannian manifolds», dans *Computer Vision - ECCV 2012*, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-642-33712-3, p. 1–14.

Karcher, H. 1977, «Riemannian center of mass and mollifier smoothing.», *Communication on Pure and Applied Mathematics*, vol. 30, p. 509–541.

Le, H. 2001, «Locating fréchet means with application to shape spaces», *Advances in Applied Probability*, vol. 33.

Pennec, X. 2013, «Bi-invariant means on lie groups with cartan-schouten connections», dans *Geometric Science of Information*, édité par F. Nielsen et F. Barbaresco, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-642-40020-9, p. 59–67.

Approximate full-conformal prediction in RKHS

Davidson Lova Razafindrakoto [1,2] & Alain Celisse [1] & Jérôme Lacaille [2]

[1] *Laboratoire SAMM, Univerisité Paris 1 Panthéon-Sorbonne, France*
[2] *Safran Aircraft Engines, France*

davidson-lova.razafindrakoto@safrangroup.com

Résumé. La prédiction "full-conformal" est un cadre qui formule implicitement des régions prédictives de confiance pour une large gamme d'estimateurs, sans hypothèse de distribution. Cependant, une limitation classique du cadre "full-conformal" est le calcul de la région prédictive de confiance, qui est en général impossible car il nécessite d'entraîner un nombre infini d'estimateurs (pour des prédictions à valeur réelle par exemple). Cette limitation est contournée en régression ridge à noyau puisqu'il existe une expression de forme close pour l'estimateur. Cependant, une telle expression n'est pas disponible en général pour des fonctions pertes non-quadratiques spécifiques à d'autres tâches, comme par exemple la perte Pinball en régression quantile. Le but principal du présent travail est de décrire une stratégie générique pour concevoir une approximation précise de la région prédictive "full-conformal" qui peut être calculée de manière efficace. En plus de cette région de confiance approchée, une quantification théorique de la précision de cette approximation est développée, en fonction de la régularité de la fonction de perte et de score. La nouvelle notion d'épaisseur est introduite pour quantifier l'écart entre la région de confiance approchée et celle "full-conformal".

Mots-clés. Apprentissage sur RKHS, Régression ridge à noyau, Région prédictive de confiance, Prédiction conforme

Abstract. full-conformal prediction is a framework that implicitly formulates distribution-free confidence predictive regions for a wide range of estimators. However, a classical limitation of the full-conformal framework is the computation of the confidence predictive regions, which is usually impossible since it requires training infinitely many estimators (for real-valued prediction for instance). This limitation is bypassed in kernel ridge regression since there exists a closed form expression for the estimator. However, such a closed form expression is not available in general for non-quadratic loss functions specific to other tasks, for instance the Pinball loss in quantile regression. The main purpose of the present work is to describe a generic strategy for designing a tight approximation to the full-conformal predictive region that can be efficiently computed. Along with this approximate confidence region, a theoretical quantification of the tightness of this approximation is developed, depending on the smoothness assumptions on the loss function. The new notion of thickness is introduced for quantifying the gap between the approximate confidence region and the full-conformal one.

Keywords. RKHS Learning, Kernel ridge regression, Confidence predictive region, Conformal prediction

1 État de l'art

Étant donné un prédicteur, la prédiction conforme (Vovk et al., 2005) est un cadre qui formule implicitement une *région prédictive de confiance* contenant la sortie non-observée à partir de ce prédicteur pour n'importe quel niveau de confiance. La première instance de ce cadre est la prédiction "full-conformal" (Vovk et al., 2022, Section 2.2.4). Cependant, la prédiction "full-conformal" nécessite d'entraîner autant de prédicteur que le cardinal de l'ensemble de prédictions possibles (Section 2.2.4 in Vovk et al. (2022)). Ceci rend le calcul de la région prédiction "full-conformal" impossible en pratique.

Ceci explique pourquoi des approches alternatives ont été conçues telles que le *split-conformal* (Papadopoulos, 2008), ou *cross-conformal* (Vovk, 2015) pour en citer quelques-uns. Contrairement au "full-conformal", les prédictions split- et cross-conformal nécessitent seulement le calcul d'un nombre fixe (petit) de prédicteurs. Le principal inconvénient de ces approximations est la perte d'information induite en considérant un sous-ensemble des observations disponibles pour l'entraînement. Cette perte d'information induit souvent des régions prédictives plus larges en pratique (voir Section 3.7 in Papadopoulos (2008) et Section 4.1 in Vovk et al. (2022)). Notons qu'il n'y a pas de quantification précise de la perte d'information induite par la prédiction spit- ou cross-conformal dans litérature à notre connaissance.

Un autre avenue de recherche exploite les configurations spécifiques où une expression de forme close est disponible pour exprimer le prédicteur entraîné en fonction de la sortie possible (voir Nouretdinov et al. (2001) pour la régression ridge, Burnaev and Nazarov (2016) pour la régression ridge à noyau, Ndiaye and Takeuchi (2019) pour Elastic net, et Lei (2019) pour LASSO). Malheureusement, de tels calculs explicites de la région "full-conformal" sont encore limités à des configurations spécifiques et restent irréalisables en toute généralité.

Contributions. Le présent travail quant à lui, décrit une stratégie *générique (sans hypothèse de distribution) pour approcher les régions full-conformal* sans découpage en amont de l'ensemble d'observations disponibles. L'idée de la conception d'une telle approximation n'est pas nouvelle. Les bornes d'uniforme stabilité sont exploitées par Ndiaye (2022) pour construire des régions prédictives "full-conformal" approchées, qui requièrent l'entraînement de seulement un prédicteur. Bien que cette approximation contient la "full-conformal", aucune garantie théorique n'est donnée sur la taille de l'approximation. De plus, l'approximation calculée souffre du comportement du pire des cas encodé par la notion de stabilité uniforme utilisée dans l'approximation.

En revanche, la principale contribution de ce travail est de présenter une instance de cette stratégie générique exploitant des hypothèses plus fortes sur la fonction de perte. La région prédictive résultante étant plus précise que celle bâtie sur la stabilité uniforme.

2 Région "full-conformal" et approximation

Supposons que $D = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ désigne un ensemble de N variables aléatoires *indépendantes et identiquement distribuées* avec une loi de probabilité P sur $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$.

Étant donnée une nouvelle $(X_{N+1}, Y_{N+1}) \sim P$ indépendante, supposons que X_{N+1} est observée. Le but de la prédiction conforme est de formuler une *région prédictive de confiance* $\hat{C}_\alpha(X_{N+1})$ bâtie sur un prédicteur $\hat{f}$ évaluer en X_{N+1} telle que, pour tout $\alpha \in (0, 1)$,

$$\mathbb{P}\left[Y_{N+1} \in \hat{C}_\alpha(X_{N+1}) \right] \geq 1 - \alpha.$$

La région (prédictive) *full-conformal* $\hat{C}_\alpha^{Full}(X_{N+1})$ (Vovk et al., 2022, Section 2.2.4) est donnée par

$$\hat{C}_\alpha^{Full}(X_{N+1}) = \{y \in \mathcal{Y} : \hat{\pi}_D(X_{N+1}, y) > \alpha\},$$

où $\hat{\pi}_D(X_{N+1}, \cdot)$ désigne la *fonction de p-valeur conforme* à l'équation (Cp-value) bâtie sur les *scores de non-conformités*. Le *score de non-conformité* du couple (X_i, Y_i) étant donné y et le prédicteur $\hat{f}_{D^y}$, est défini par

$$\begin{aligned} S_i(y) &:= S(Y_i, \hat{f}_{D^y}(X_i)), \qquad \text{if } 1 \leq i \leq N \\ S_{N+1}(y) &:= S(Y_i, \hat{f}_{D^y}(X_i)), \qquad \text{if } i = N+1. \end{aligned} \tag{NCScores}$$

où $S(\cdot, \cdot) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ une fonction qui désigne une *mesure de non-conformité* (Balasubramanian et al., 2014, Définition 1.3), et $\hat{f}_{D^y}$ un prédicteur entraîné sur $D^y := \{(X_1, Y_1), \ldots, (X_{N+1}, y)\}$, pour tout $y \in \mathcal{Y}$. La fonction de p-valeur conforme résultante $\hat{\pi}_D(X_{N+1}, \cdot)$ est donnée par

$$y \in \mathcal{Y} \mapsto \hat{\pi}_D(X_{N+1}, y) := \frac{1 + \sum_{i=1}^N \mathbb{1}_{\{S_i(y) \geq S_{N+1}(y)\}}}{N+1}. \tag{Cp-value}$$

Définition 1 *(Région préditive "full-conformal" approchée) Pour tout niveau de contrôle de confiance $\alpha \in (0, 1)$, supposons que pour chaque $1 \leq i \leq N+1$, et $\forall y \in \mathcal{Y}$, $\tilde{S}_i(y)$ est une approximation du* score de non-conformité *$S_i(y)$ et qu'il existe une borne supérieure sur la qualité de l'approximation $0 \leq \tau_i(y) < +\infty$ telle que $|S_i(y) - \tilde{S}_i(y)| \leq \tau_i(y)$. Alors la p-valeur "full-conformal" approchée $\tilde{\pi}_D(X_{N+1}, \cdot)$ est donnée par*

$$y \in \mathcal{Y} \mapsto \tilde{\pi}_D(X_{N+1}, y) := \frac{1 + \sum_{i=1}^N \mathbb{1}_{\left\{\tilde{S}_i(y) + \tau_i(y) \geq \tilde{S}_{N+1}(y) - \tau_{N+1}(y)\right\}}}{N+1}. \tag{1}$$

De plus, la région prédictive qui en résulte est donnée par

$$\tilde{C}_\alpha(X_{N+1}) := \{y \in \mathcal{Y}, \tilde{\pi}_D(X_{N+1}, y) > \alpha\},$$

et par construction, elle contient la région "full-conformal", c'est-à-dire

$$\hat{C}_\alpha^{Full}(X_{N+1}) \subseteq \tilde{C}_\alpha(X_{N+1}).$$

3 Qualité d'approximation et une instance

La notion d'*épaisseur* étant la mesure de Lebesgue de la différence symétrique entre la région approchante et la région "full-conformal" est proposée comme mesure de la qualité d'approximation. Similairement à la région approchante $\tilde{C}_\alpha(X_{N+1})$ étant une approximation supérieure au sens de l'inclusion, une région approchante inférieure est construite en changeant les signes devant les termes $\tau_i(\cdot)$s dans l'Équation 1. Avec cela, une borne supérieure sur l'*épaisseur* est construite en prenant la mesure de Lebesgue de la différence symétrique entre la région approchante supérieure et celle inférieure.

Proposition 1 *Pour tout niveau de contrôle de confiance $\alpha \in (0,1)$, il existe une fonction* $\mathrm{GapR}_D(X_{N+1})$ *entièrement bâties sur $\tilde{S}_i(\cdot)$s et $\tau_i(\cdot)$s telles que*

$$\mathrm{THK}_\alpha(X_{N+1}) := \mathcal{L}\left(\tilde{C}_\alpha(X_{N+1}) \Delta \hat{C}_\alpha(X_{N+1})\right) \leq \mathrm{GapR}_\alpha(X_{N+1}). \tag{2}$$

Dans la suite, le prédicteur est choisi parmi les fonctions contenu dans un espace d'Hilbert à noyau reproduisant (RKHS) $\mathcal{H} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$, avec $\kappa_{\mathcal{H}}(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ le noyau correspondant et $\|\cdot\|_{\mathcal{H}} : \mathcal{H} \to \mathbb{R}_+$ la norme correspondante. Étant donné un ensemble d'observations D, le predicteur $\hat{f}_D$ est celui qui minimise la fonction de risque empirique régularisée, c'est-à-dire pour $\lambda > 0$

$$\hat{f}_D \in \operatorname*{argmin}_{f \in \mathcal{H}} \left\{ \frac{1}{\mathrm{Card}(D)} \sum_{(x,y)\in D} \ell\left(y, f(x)\right) + \lambda \|f\|_{\mathcal{H}}^2 \right\},$$

où $\ell(\cdot,\cdot) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ est une fonction de perte fixée selon la tâche. Ce formalisme permet de modéliser des relations non-linéaires entre l'entrée et la prédiction, éviter le sur-apprentissage avec la pénalisation ridge, et prédire des sorties spécifiques à la tâche imposée par la fonction de perte. Avec cette classe de prédicteur, les scores approchées sont calculées en remplaçant le prédicteur $\hat{f}_{D^y}$ dans la définition (NCScores) par $\hat{f}_{D^z}$ pour un $z \in \mathcal{Y}$ fixé. Ceci réduit le nombre le prédicteur utilisé à un seul au lieu de $\mathrm{Card}(\mathcal{Y})$, tout en gardant le nombre de scores (pas de découpages). La proposition suivante décrit deux bornes supérieures sur la qualité d'approximation des scores : la première basée sur la stabilité uniforme, le deuxième exploitant davantages de régularité de la fonction perte.

Proposition 2 *S'il existe $\rho > 0$ et $\gamma > 0$ telles que la fonction de perte $\ell(\cdot,\cdot)$ est convexe ρ-Lipschitz et la fonction de score $S(\cdot,\cdot)$ est γ-Lipschitz alors $\forall 1 \leq i \leq N+1$*

$$\left|S_i(y) - \tilde{S}_i(y)\right| = (K_i)^{1/2} (K_{N+1})^{1/2} \frac{\gamma\rho}{\lambda(N+1)} =: \tau_i^{(0)} \tag{3}$$

En revanche si $\ell(\cdot,\cdot)$ est convexe et deux fois continûment dérivable alors

$$\left|S_i(y) - \tilde{S}_i(y)\right| = (K_i)^{1/2} (K_{N+1})^{1/2} \frac{\gamma\rho^{(1)}(y)}{\lambda(N+1)} =: \tau_i^{(1)}(y) \tag{4}$$

où $\rho^{(1)}(y) = \frac{1}{2}\left|\ell_{\hat{y}}(y, \hat{f}_{D^z}(X_{N+1})) - \ell_{\hat{y}}(z, \hat{f}_{D^z} X_{N+1})\right|$ avec $\ell_{\hat{y}}(\cdot,\cdot)$ la dérivée partielle de $\ell(\cdot,\cdot)$ selon son second argument et $K_i := \kappa_{\mathcal{H}}(X_i, X_i)$.

Notons que la différence entre les deux bornes est au niveau de la constante de Lipschitz ρ qui désigne le pire des cas en terme de dérivé première de la fonction de perte, remplacée par $\rho^{(1)}(y)$ qui lui utilise la dérivé première. La proposition suivante compare les bornes supérieures ainsi que les épaisseurs et GapRs résultantes.

Proposition 3 *S'il existe $\rho > 0$ et $\gamma > 0$ telles que la fonction de perte $\ell(\cdot, \cdot)$ est convexe ρ-Lipschitz et deux fois continûment dérivable, et la fonction de score $S(\cdot, \cdot)$ est γ-Lipschitz alors $\forall 1 \leq i \leq N+1, \forall y \in \mathcal{Y}, 0 \leq \tau_i^{(1)}(y) \leq \tau_i^{(0)}$ ce qui implique $\forall \alpha \in (0,1)$*

$$\mathrm{THK}_\alpha^{(1)}(X_{N+1}) \leq \mathrm{THK}_\alpha^{(0)}(X_{N+1}),$$
$$\mathrm{GapR}_\alpha^{(1)}(X_{N+1}) \leq \mathrm{GapR}_\alpha^{(0)}(X_{N+1}).$$

Ceci indique que les bornes non-uniformes sont plus étroites que les bornes uniformes et les régions induites sont plus précises. Ceci est illustrée dans la figure suivante sur des données simulées en tirant un ensemble d'observations D de taille N, à partir duquel est évaluée les GapRs. Le script est disponible ici.

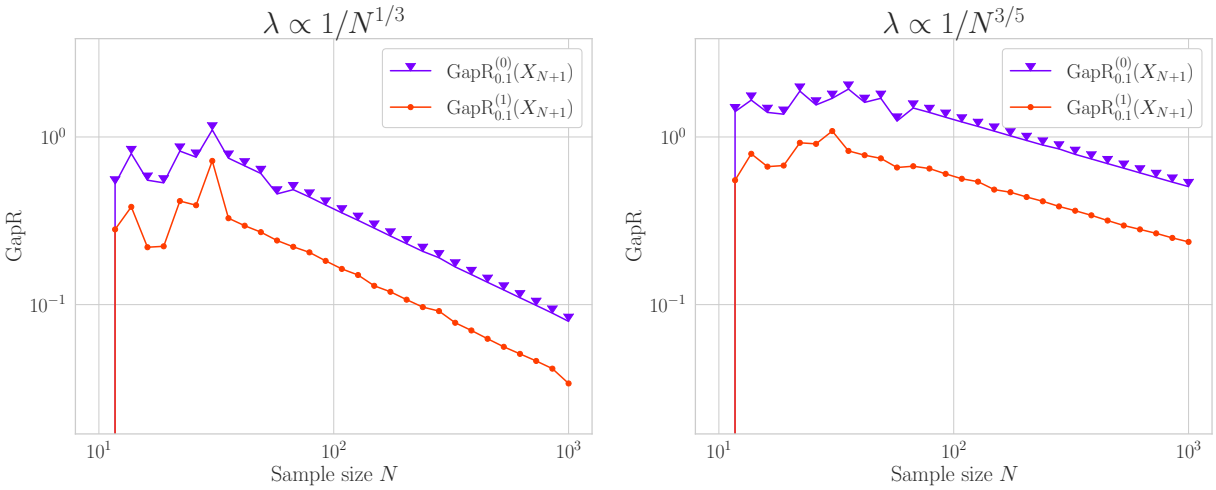


Figure 1: Comparaison de l'évolution de $\mathrm{GapR}_{0.1}^{(0)}(X_{N+1})$ et $\mathrm{GapR}_{0.1}^{(1)}(X_{N+1})$, en fonction de la taille N de l'ensemble d'observations (en échelle log log). Ceci pour deux vitesses de décroissance pour λ : $\frac{0.5}{N^{1/3}}$ à gauche et $\frac{0.5}{N^{3/5}}$ à droite. En prenant, $z = 0$, $\ell(\cdot, \cdot)$ la perte 0.5-Pinball 0.5-lissée, $k_{\mathcal{H}}(\cdot, \cdot)$ le noyau 0.1-Laplacien, $S(\cdot, \cdot)$ la déviation absolue et le jeu de donnée synthétique est make_regression(N, p = 100, noise = 1) de sklearn.

Dans ce cas, les hypothèses de la proposition 3 sont vérifiées car la perte 0.5-Pinball 0.5-lissée est convexe 0.5-Lipschitz et deux fois dérivable, et la déviation absolue est 1-Lipschitz. Remarquons que GapR diminue avec une pente qui varie en fonction de celle de λ. Comme théorisé, $\mathrm{GapR}_{0.1}^{(1)}(X_{N+1})$ est plus bas que $\mathrm{GapR}_{0.1}^{(0)}(X_{N+1})$. Ceci indique qu'au moins dans le pire des cas, les régions approchées proposées sont plus précises que les régions "stable conformal".

4 Discussion

Seulement une instance de région approchée est présentée. Cependant, les bornes peuvent être d'autant plus améliorer quand la fonction de perte est fortement convexe, et spécifier pour d'autres profils de fonction de score, par exemple C^2 et β-smooth. De plus, des prédicteurs approchés plus précis peuvent être aussi bâtis en utilisant des fonctions d'influences, les régions résultantes d'autant plus précises. Parallèlement à cela, les prochains travaux viseront à instancier le schéma présenté ici pour d'autres classes prédicteurs, par exemple ceux qui sont bâtis sur des algorithmes itératifs de recherche de minimiseur telle que la descente de gradient. Ceci sera suivi d'une étude théorique sur la qualité de la région full-conformal elle-même et de ses approximations, vis-à-vis d'une région prédictive oracle qui connaît la distribution des observations.

Bibliographie

Balasubramanian, V., Ho, S.-S., and Vovk, V. (2014). *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes.

Burnaev, E. and Nazarov, I. (2016). Conformalized kernel ridge regression. In *2016 15th IEEE international conference on machine learning and applications (ICMLA)*, pages 45–52. IEEE.

Lei, J. (2019). Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106(4):749–764.

Ndiaye, E. (2022). Stable conformal prediction sets. In *International Conference on Machine Learning*, pages 16462–16479. PMLR.

Ndiaye, E. and Takeuchi, I. (2019). Computing full conformal prediction set with approximate homotopy. *Advances in Neural Information Processing Systems*, 32.

Nouretdinov, I., Melluish, T., and Vovk, V. (2001). Ridge regression confidence machine. In *ICML*, pages 385–392. Citeseer.

Papadopoulos, H. (2008). Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer.

Vovk, V. (2015). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Vovk, V., Gammerman, A., and Shafer, G. (2022). *Algorithmic Learning in a Random World*. Second edition.

Sélection de variables

Estimation et sélection de variables dans les modèles non linéaires à effets mixtes.

Antoine Caillebotte [1,2] & Estelle Kuhn [2] & Sarah Lemler [3]

[1] *Université Paris-Saclay, INRAE, UMR GQE-Moulon, France, caillebotte.antoine@inrae.fr,*
[2] *Université Paris-Saclay, INRAE, UR MaIAGE, France, estelle.kuhn@inrae.fr,*
[3] *Université Paris-Saclay, CentraleSupélec, Laboratoire MICS, France, sarah.lemler@centralesupelec.fr*

Résumé.

Nous considérons des modèles non linéaires à effets mixtes incluant des covariables de grande dimension pour modéliser les paramètres individuels. L'objectif visé est d'identifier les covariables pertinentes et d'estimer les paramètres du modèle. Nous combinons un estimateur pénalisé de type LASSO à un critère eBIC de choix de modèle pour sélectionner les covariables d'intérêt. Puis nous estimons les paramètres par maximum de vraisemblance dans le modèle réduit. Nous calculons l'estimateur pénalisé de type LASSO par un algorithme de descente de gradient proximal pondéré avec un pas adaptatif vectoriel de type Adagrad. Ce choix d'algorithme nous permet en particulier de considérer des modèles n'appartenant pas nécessairement à la famille exponentielle. Nous illustrons les performances de la méthode dans un modèle non linéaire de croissance logistique à effets mixtes.

Mots-clés. Modèle non linéaire à effets mixtes, grande dimension, sélection de variables, pénalité LASSO, descente de gradient stochastique, proximal pondéré.

Abstract.

We consider nonlinear mixed effects models including high-dimensional covariates to model individual parameters. The objective is to identify relevant covariates and estimate model parameters. We combine a penalized LASSO-type estimator with an eBIC model choice criterion to select the covariates of interest. Then we estimate the parameters by maximum likelihood in the reduced model. We calculate the LASSO-type penalized estimator by a weighted proximal gradient descent algorithm with an Adagrad-type vector adaptive step. This choice allows us in particular to consider models that do not necessarily belong to the curved exponential family. We illustrate the performance of the method in a nonlinear mixed-effects logistic growth model.

Keywords. nonlinear mixed effects model, high dimension, variable selection, LASSO penalty, stochastic gradient descent, weighted proximal.

1 Introduction

Les modèles à effets mixtes sont des outils de modélisation statistique bien adaptés pour analyser des données avec des structures hiérarchiques et des mesures répétées dans de nombreux domaines appliqués tels que l'agronomie, la pharmacologie et même l'économie Pinheiro et Bates [2006]. Ils sont composés de deux niveaux de modélisation : d'une part une modélisation structurelle du phénomène d'intérêt paramétrée pour chaque individu de la population par des paramètres individuels spécifiques, d'autre part une modélisation de ces paramètres individuels sous forme de variables aléatoires décrivant leur variabilité au sein de la population.

La modélisation structurelle du phénomène d'intérêt peut être linéaire ou non linéaire en les paramètres individuels. Les modélisations de type non linéaire permettent notamment rendre compte de phénomènes complexes via des modèles mécanistes intégrant des connaissances physiques ou biologiques. Les paramètres de tels modèles présentent souvent un grand intérêt pratique car ils sont interprétables d'un point de vue appliqué. Dans un tel contexte, un objectif central est de caractériser et d'expliquer la variation de ces paramètres individuels au sein de la population. Ainsi la modélisation des paramètres individuels combine des covariables descriptives des individus et des effets aléatoires au niveau individuel.

Dans de nombreux contextes, ces covariables descriptives peuvent être de grande dimension. Du point de vue de l'inférence, l'objectif est alors d'identifier un petit nombre d'entre elles qui sont les plus pertinentes pour expliquer les variabilités observées au sein de la population. et d'estimer les paramètres dans le modèle réduit. La sélection des covariables pertinentes se fait classiquement via une approche de régularisation. Dans le cadre des modèles à effets mixtes, une difficulté supplémentaire apparaît du fait de la présence d'effets aléatoires non observés. Il s'agit d'un contexte classique de modèles à variables latentes. L'inférence de paramètres peut se faire par exemple par maximum de vraisemblance via des algorithmes Expectation Maximisation (EM) Delyon *et al.* [1999] ou de type gradient stochastique Baey *et al.* [2023]. Dans le cadre de modèles appartenant à la famille exponentielle, les algorithmes de type EM sont faciles à mettre en œuvre et possèdent de bonnes propriétés théoriques. En revanche, à notre connaissance, il n'existe pas de résultats théoriques de convergence en dehors du cadre. De plus, la mise en œuvre de ces algorithmes en dehors de la famille exponentielle est plus complexe. Les méthodes de gradient stochastique peuvent être appliquées dans des modèles plus génériques, notamment en dehors de la famille exponentielle. Des garanties théoriques ont été établies. Cependant, ces méthodes sont en pratique très sensibles au choix du pas du gradient, en particulier dans les espaces de paramètres de grande dimension en raison de l'hétérogénéité des différentes composantes du gradient. Des choix adaptatifs du pas de gradient Duchi *et al.* [2011] et des procédures basées sur le préconditionnement du gradient Baey *et al.* [2023] ont été proposés.

Plusieurs approches de régularisation ont été développées avec une pénalité LASSO dans le contexte de modèles linéaires et linéaires généralisés à effets mixtes Schelldorfer *et al.* [2014]. Dans le cas de modèles non linéaires à effets mixtes appartenant à la famille exponentielle, un estimateur avec des pénalités plus générales Fort *et al.* [2019] et une approche bayésienne basée sur des distributions Spike and Slab ont été développées par Naveau *et al.* [2024].

Dans cette contribution, nous proposons une méthode pour calculer un estimateur du maximum de vraisemblance régularisé via une pénalité de type LASSO dans un modèle à effets mixtes général.

2 Modèle Non Linéaire à Effets Mixtes

On considère une population de N individus et pour chaque individu J mesures répétées. On note $Y_{i,j}$ la $j-$ième observation du $i-$ième individu pour $1 \leq i \leq N$ et $1 \leq j \leq J$ à valeur dans $\mathbb{R}^d$ que nous modélisons par un modèle à effets mixtes non linéaire Pinheiro et Bates [2006] :

$$\begin{cases} Y_{i,j} & = h(\alpha, \tau_{i,j}, Z_i) + \varepsilon_{i,j} \\ Z_i & \overset{i.i.d}{\sim} \mathcal{N}(X_i\beta, \Omega) \\ \varepsilon_{i,j} & \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma) \end{cases}$$

où h est une fonction non linéaire dépendant du paramètre de population $\alpha \in \mathbb{R}^a$, des covariables individuelles $\tau_{i,j}$ et des paramètres individuels Z_i. La formulation proposée très générale permet de considérer h comme fonction à la fois des paramètres de population et des paramètres individuels.

Le bruit $\varepsilon_{i,j}$ est supposé gaussien centré de covariance inconnue Σ. Les paramètres individuels Z_i sont des variables aléatoires gaussiennes d'espérance $X_i\beta$ et de matrice de covariance Ω, où X_i est une matrice de covariables et $\beta \in \mathbb{R}^p$. On suppose que (Z_i) et $(\varepsilon_{i,j})$ sont indépendants. Les paramètres inconnus du modèle sont $\theta = (\alpha, \beta, \Omega, \Sigma) \in \Theta$. Nous considérons le cas où p peut être beaucoup plus grand que N.

3 Inférence des paramètres et sélection de variables

3.1 Méthodologie

Dans le contexte des modèles à variables latentes, la vraisemblance marginale, notée g, est obtenue en intégrant la vraisemblance complète, notée f, selon les variables latentes :

$$g(\theta; Y) \quad = \quad \prod_{i=1}^{N} \int f(Y_i, Z_i; \theta) dZ_i$$

Dans notre contexte, nous introduisons une régularisation pour gérer la dimension élevée des covariables, et considérons un estimateur du maximum de vraisemblance pénalisé par la procédure LASSO Tibshirani [1996].

$$\hat{\theta}_\lambda^{LASSO} = \arg\max_{\theta \in \Theta} \left\{ \log g(\theta; Y) - \text{pen}_\lambda(\theta) \right\}. \tag{1}$$

où $\lambda > 0$. Plus la valeur de λ est grande, plus θ sera contraint d'avoir des composantes à zéro. Dans notre contexte, nous considérons une pénalité LASSO appliquée uniquement au paramètre β. Nous choisissons la valeur de ce paramètre en minimisant le critère d'information bayésien étendu noté eBIC Chen et Chen [2008] qui intègre une pénalité pour les modèles de grande dimension défini par

$$\text{eBIC}(\lambda) = -2\log g_\lambda(\hat{\theta}_\lambda^{MLE}; Y) + |\hat{S}_\lambda| \log(N(J+1)) + 2\log\left(\binom{p}{|\hat{S}_\lambda|}\right) \tag{2}$$

Nous considérons une grille Λ et pour chaque valeur $\lambda \in \Lambda$ nous sélectionnons le support $\hat{S}_\lambda = \{j \in \{1,\dots,p\}, \hat{\beta}_j^{LASSO} \neq 0\}$ du vecteur β en calculant l'estimateur pénalisé (1) puis nous calculons l'estimateur du maximum de vraisemblance du paramètre dans le modèle réduit à ce support, noté $\hat{\theta}_\lambda^{\text{MLE}}$. Nous calculons ensuite pour tout $\lambda \in \Lambda$ le critère eBIC(λ). Pour finir nous sélectionnons la valeur $\hat{\lambda}$ qui minimise le critère eBIC(λ) et considérons l'estimateur $\hat{\theta}_{\hat{\lambda}}^{\text{MLE}}$.

3.2 Algorithme numérique pour calculer l'estimateur pénalisé

Nous utilisons une descente de gradient stochastique préconditionnée adapté aux modèles à variables latentes généraux Baey *et al.* [2023]. L'algorithme 1 est divisé en trois étapes : simulation des variables latentes via un échantillonneur de Metropolis-Hastings, descente de gradient préconditionnée par la matrice diagonale déduite de Adagrad Duchi *et al.* [2011] sur la log-vraisemblance complète, calcul du proximal pondéré associé Fort *et al.* [2019]; Chouzenoux *et al.* [2014].

Algorithm 1: Gradient Proximal Stochastique avec Adagrad (SPGD-Ada)

Require: Nombre d'itérations $K \geq 1$; séquence de pas $(\gamma_k)_{k\geq 1}$
Initialize Point de départ $\theta_0 \in \mathbb{R}^d$, Δ_0, $s_0 = 0$, γ_0
for $k = 1$ **to** K **do**
  • **Étape de simulation :**
    Obtenir $Z^{(k)}$**en utilisant une étape de Metropolis-Hastings**
  • **Calcul du gradient :** $v_k = \frac{1}{N}\sum_{i=1}^{N} \nabla \log f_{\theta_k}(Y_i, Z_i^{(k)})$
  • **Calcul du pas adaptatif :** $s_{k,l} = s_{k-1,l} + v_{k,l}^2$ **pour** $l = 1, ..., p$
  • **Descente de gradient :**
    • **Étape Forward :** $\omega_{k,l} = \theta_{k-1,l} + \gamma_0 v_{k,l}/\sqrt{s_{k,l}}$ **pour** $l = 1, ..., p$
    • **Étape Backward :** $\theta_k = \text{Prox}_{\gamma \text{pen}_\lambda}(\omega_k)$
end
return $\hat{\theta} = \theta_K$

4 Étude numérique de simulation

Nous avons tout d'abord comparé les performances obtenues par les méthodes SPGD-Ada et GLMMLASSO Groll et Tutz [2014] sur un modèle linéaire à effets mixtes.Les résultats obtenus sont d'ordre comparable et seront détaillés en exposé.

Nous considérons ensuite un modèle non linéaire de croissance logistique et incluons les variables de grande dimension sur le paramètre du temps de demi vie :

$$\begin{cases} Y_{i,j} & = \frac{Z_{i1}}{1+\exp\left(-\frac{\tau_{ij}-Z_{i2}}{\alpha}\right)} + \varepsilon_{i,j} \\ Z_{i1} & \overset{i.i.d}{\sim} \mathcal{N}(\mu_1, \nu_1^2) \\ Z_{i2} & \overset{i.i.d}{\sim} \mathcal{N}(\mu_2 + X_i\beta, \nu_2^2) \\ \varepsilon_{i,j} & \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma) \end{cases} \tag{3}$$

Nous générons 100 jeux de données selon (3), avec $N = 100$ et $J = 15$. pour $\mu = (200, 1200)$, $(\nu_1^2, \nu_2^2) = (7^2, 30^2)$, $\alpha = 300$ et $\Sigma^2 = 30$. Le vecteur β est choisi tel que les trois premières composantes sont égales à $(100, 200, -300)$ et les autres sont nulles pour $p \in \{100, 200, 500\}$. Les covariables $X_{i,k}$ suivent une loi uniforme sur $[-1, 1]$. Nous mettons en oeuvre la méthode proposée. Les résultats obtenus sont de très bonne qualité et détaillés dans les tables 1 et 2.

Table 1: Estimation et erreur quadratique moyenne relative (RRMSE) pour 100 répétitions.

		$P = 100$		$P = 200$		$P = 500$	
	θ^*	$\hat{\theta}$	RRMSE	$\hat{\theta}$	RRMSE	$\hat{\theta}$	RRMSE
μ_1	200	199.9	0.40	199.9	0.40	199.9	0.39
μ_2	1200	1200.8	0.29	1201.0	0.29	1200.8	0.30
$\gamma_{1,1}^2$	49	48.0	6.79	47.9	6.81	48.0	6.84
$\gamma_{2,2}^2$	900	809.8	20.61	774.0	23.25	722.7	41.77
α	300	300.2	0.58	300.3	0.54	300.1	0.55
σ^2	30	30.0	3.82	30.1	3.73	30.1	3.87
β_1	100	99.8	5.78	100.3	7.10	97.5	15.89
β_2	200	199.9	3.39	199.2	3.37	199.2	3.44

Table 2: Moyenne pour la sensibilité (Se), spécificité (Sp), précision (Ac) pour 100 répétitions.

	$P = 100$	$P = 200$	$P = 500$
Ac	0.970	0.973	0.977
Se	1.000	1.000	0.993
Sp	0.969	0.973	0.977

5 Discussion

Dans ce travail, nous avons considéré un modèle non linéaire à effets mixtes où les paramètres individuels sont définis à partir de covariables de grande dimension et d'effets aléatoires. Nous proposons une méthode numérique de sélection de variables basée sur une pénalité LASSO et un critère eBIC. Une perspective intéressante consiste à étudier les propriétés de l'estimateur et de la prédiction après l'étape du choix de modèle.

Remerciements. Ce travail a été financé par le projet Stat4Plant ANR-20-CE45-0012.

References

Baey, C., Delattre, M., Kuhn, E., Leger, J.-B. et Lemler, S. (2023). Efficient preconditioned stochastic gradient descent for estimation in latent variable models. *In Proceedings of the 40th International Conference on Machine Learning*, ICML'23.

Chen, J. et Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chouzenoux, E., Pesquet, J.-C. et Repetti, A. (2014). Variable metric forward—backward algorithm for minimizing the sum of a differentiable function and a convex function. *J. Optim. Theory Appl.*, 162(1):107–132.

Delyon, B., Lavielle, M. et Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pages 94–128.

Duchi, J., Hazan, E. et Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159.

Fort, G., Ollier, E. et Samson, A. (2019). Stochastic proximal-gradient algorithms for penalized mixed models. *Statistics and Computing*, 29:231–253.

Groll, A. et Tutz, G. (2014). Variable selection for generalized linear mixed models by l1-penalized estimation. *Statistics and Computing*, 24(2):137–154.

Naveau, M., Kon Kam King, G., Rincent, R., Sansonnet, L. et Delattre, M. (2024). Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the saem algorithm. *Statistics and Computing*, 34(1):53.

Pinheiro, J. et Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer science & business media.

Schelldorfer, J., Meier, L. et Bühlmann, P. (2014). Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using ℓ_1-penalization. *Journal of Computational and Graphical Statistics*, 23(2):460–477.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Sélection de variables par ACP supervisée : réduction des faux positifs en grande dimension

Insha Ullah[1] & Kerrie Mengersen[2] & Anthony N Pettitt[2] & Benoit Liquet[3]

[1] *ANU College of Business and Economics, The Australian National University, Australia,*
[2] *School of Mathematical Sciences, Queensland University of Technology, Austrlia*
[3] *Université de Pau et des Pays de l'Adour, E2S-UPPA, PAU, France,*
benoit.liquet@univ-pau.fr

Résumé. Les jeux de données en grande dimension, où le nombre de variables p excède largement le nombre d'observations n, posent des défis majeurs en classification statistique en raison des problèmes de surajustement. Nous présentons une méthode innovante de sélection de variables fondée sur l'analyse en composantes principales (ACP), spécialement conçue pour les problèmes de classification binaire et les études cas-témoins. Cette approche combine efficacité computationnelle et robustesse statistique, s'adaptant particulièrement aux scénarios où $p \gg n$. Par des simulations approfondies et des applications à des données génomiques complexes - incluant des études d'expression génique de leucémie aiguë lymphoblastique infantile et une analyse GWAS de variants associés à des traits morphologiques du riz - nous démontrons la supériorité de notre méthode. Celle-ci permet non seulement une identification précise des variables pertinentes, mais aussi un contrôle rigoureux du taux de faux positifs (FDR) grâce à une modélisation explicite des corrélations entre variables et un prétraitement efficace des données. Son implémentation légère, comparable en rapidité aux méthodes univariées tout en offrant la puissance des approches multivariées, en fait un outil particulièrement adapté aux défis actuels de l'analyse des données omiques.

Mots-clés. Mot-clé classification binaire, étude cas-témoins, sélection de variables, données d'expression génique, GWAS

Abstract. High-dimensional datasets, where the number of variables p is much larger than the number of samples n, are ubiquitous and often render standard classification techniques unreliable due to overfitting. An important research problem is feature selection, which ranks candidate variables based on their relevance to the outcome variable and retains those that satisfy a chosen criterion. This presentation proposes a computationally efficient variable selection method based on principal component analysis tailored to a binary classification problem or case-control study. This method is accessible and is suitable for the analysis of high-dimensional datasets. We demonstrate the superior performance of our method through extensive simulations. A semi-real gene expression dataset, a challenging childhood acute lymphoblastic leukemia gene expression study, and a GWAS that attempts to identify single-nucleotide polymorphisms (SNPs) associated with rice grain length further demonstrate the usefulness of our method in genomic applications. We expect our method to accurately identify important features but at the same time reduce the False Discovery Rate (FDR) by accounting for the correlation between variables and by de-noising data in the training phase, which also makes it robust to mild outliers in the training data. As it is almost as fast as univariate filters, our method allows valid statistical inference.

Keywords. binary classification, case-control study, feature selection, gene-expression data, GWAS,

1 Introduction

L'analyse des données en grande dimension, caractérisées par un nombre de variables p bien supérieur au nombre d'observations n, pose des défis méthodologiques majeurs. Ces défis sont particulièrement aigus dans les domaines comme la génomique, où des milliers de caractéristiques sont mesurées sur un nombre limité d'échantillons. Dans ce contexte, les méthodes statistiques traditionnelles deviennent inopérantes ou produisent des modèles surajustés, performants sur les données d'entraînement mais médiocres en généralisation.

Face à ce constat, la sélection de variables représente un enjeu crucial. Les approches existantes se divisent principalement en deux catégories : les méthodes univariées, simples et rapides mais ignorant les dépendances entre variables, et les méthodes multivariées plus complètes mais souvent coûteuses en calcul. Les méthodes hybrides tentent de combiner leurs avantages, mais peinent encore à concilier efficacité statistique et faisabilité computationnelle.

L'Analyse en Composantes Principales (ACP), bien qu'ayant fait ses preuves en réduction de dimension, présente des limites pour la sélection de variables interprétables. Les adaptations existantes, comme l'ACP sparse, restent souvent arbitraires ou inadaptées aux très grande dimensions. C'est dans ce contexte que nous proposons VSPCA (Variable Selection based on PCA), une nouvelle approche exploitant les erreurs de reconstruction entre classes pour identifier les variables discriminantes.

Après avoir exposé les fondements théoriques de notre approche, nous en démontrerons la performance à travers des études de simulation et des applications à des données génomiques réelles, en comparaison avec les méthodes existantes. Les résultats montrent que VSPCA surpasse souvent les alternatives tant en puissance de détection qu'en stabilité des sélections.

2 Méthodologie

L'Analyse en Composantes Principales (ACP) est traditionnellement utilisée comme technique de réduction de dimensionnalité et constitue un outil précieux pour visualiser des données de grande dimension dans un espace de dimension réduite. Elle transforme orthogonalement un ensemble de p variables (potentiellement corrélées) en un ensemble de p variables non corrélées appelées composantes principales (CP). Ces CP sont des combinaisons linéaires des variables originales qui maximisent successivement la variance : la première CP correspond à l'axe présentant la plus grande variance dans les données, la seconde CP à l'axe orthogonal présentant la deuxième plus grande variance, et ainsi de suite. Ainsi, les p dimensions orthogonales de variabilité sont capturées dans p CP, avec une variance cumulée égale à la variation totale des données. L'objectif principal de l'ACP étant de capturer un

maximum de variation dans les premières CP, il est courant que les premières q CP ($q \ll p$) contiennent l'information utile tandis que les autres reflètent principalement du bruit (Joliffe (2003)).

Formellement, soit y_{ij} l'observation de la j-ème variable pour le i-ème sujet ($i = 1, \ldots, n$; $j = 1, \ldots, p$). Les données sont organisées dans une matrice Y de taille $n \times p$. Après standardisation des colonnes de Y (moyenne nulle et écart-type unitaire), on obtient la matrice X :

$$x_{ij} = \frac{y_{ij} - \bar{y}_j}{s_j}.$$

En pratique, l'ACP est réalisée par décomposition en valeurs singulières (SVD) de X :

$$X = U\Gamma V^T \tag{1}$$

où V contient les vecteurs propres de la matrice de corrélation $R = \frac{1}{n-1}X^T X$. Les composantes principales sont obtenues par :

$$Z = XV \tag{2}$$

Les premières q CP ($\tilde{Z} = X\tilde{V}$ où $\tilde{V}$ contient les q premières colonnes de V) fournissent une représentation de dimension réduite. Une approximation de bas rang de X est donnée par :

$$\tilde{X} = \tilde{Z}\tilde{V}^T = X\tilde{V}\tilde{V}^T \tag{3}$$

qui constitue la meilleure approximation au sens des moindres carrés (Eckart (1936)). Le choix de q repose généralement sur un critère de variance expliquée (par exemple $\phi = 0.8$) ou sur l'examen du scree-plot.

2.1 Adaptation pour la sélection de variables (VSPCA)

L'ACP se révèle particulièrement efficace pour traiter des jeux de données de grande dimension. Dans cette étude, nous proposons une adaptation innovante de l'ACP visant à la sélection de variables pour des problèmes de classification, dépassant ainsi son usage traditionnel limité à la réduction de dimension et à la visualisation. Pour ce faire, nous exploitons les erreurs de reconstruction obtenues lors de l'approximation des observations.

Plus précisément, considérons d'abord l'erreur de reconstruction d'une observation x_{ij} définie par :

$$\hat{e}_{ij} = x_{ij} - \tilde{x}_{ij}, \tag{4}$$

où $\tilde{x}_{ij}$ représente la valeur reconstruite. Ces erreurs sont ensuite stockées dans une matrice résiduelle de taille $n \times p$:

$$\hat{E} = X - \tilde{X} = X(I - \tilde{V}\tilde{V}^T). \tag{5}$$

Dans le cadre d'une étude cas-témoins, cette approche prend tout son sens. D'une part, nous avons la matrice Y^- ($n_1 \times p$) contenant les données des témoins, et d'autre part la matrice

Y^+ $(n_2 \times p)$ regroupant les observations des cas. La première étape consiste à standardiser les données des témoins pour obtenir X^- selon :

$$x_{ij}^- = (y_{ij}^- - \bar{y}_j^-)/s_j^-.$$

Passons maintenant à la phase de modélisation. Après avoir effectué la décomposition SVD de X^-, nous obtenons $\tilde{V}^-$, contenant les q premiers vecteurs singuliers droits. Les données des cas Y^+ sont alors standardisées en utilisant les paramètres estimés sur les témoins :

$$x_{ij}^+ = (y_{ij}^+ - \bar{y}_j^-)/s_j^-.$$

Vient ensuite l'étape cruciale de reconstruction. En utilisant les composantes principales issues des témoins, nous reconstruisons approximativement les données des cas :

$$\tilde{X}^+ = X^+\tilde{V}^-\tilde{V}^{-T}.$$

Les erreurs de reconstruction qui en résultent :

$$\hat{e}_{ij}^+ = x_{ij}^+ - \tilde{x}_{ij}^+, \tag{6}$$

s'avèrent particulièrement informatives. En effet, elles sont stockées dans une matrice résiduelle spécifique :

$$\hat{E}^+ = X^+ - \tilde{X}^+ = X^+(I - \tilde{V}^-\tilde{V}^{-T}), \tag{7}$$

où l'amplitude de ces erreurs reflète directement le pouvoir discriminant des variables.

Poursuivons avec l'analyse statistique proprement dite. Pour chaque variable j, nous calculons une statistique synthétique :

$$t_j^+ = \frac{1}{n_2}\sum_{i=1}^{n_2} \hat{e}_{ij}^+, \tag{8}$$

qui, sous l'hypothèse que les données cas et témoins proviennent du même modèle, suit asymptotiquement une loi normale :

$$\sqrt{n_2}t_j^+ \xrightarrow{d} N(0, \sigma_j^2). \tag{9}$$

Nous supposons une variance commune σ^2 pour tout j afin de stabiliser les estimations de variance et afin de renforcer la robustesse de notre approche, nous estimons σ à l'aide de l'écart absolu médian (MAD) :

$$\hat{\sigma} = 1.4826\sqrt{n_2}\ \text{Médiane}\{|t_j^+ - \text{Médiane}(t_j^+)|\}, \tag{10}$$

ce qui nous permet, via le théorème de Slutsky, d'obtenir :

$$\frac{t_j^+}{\hat{\sigma}/\sqrt{n_2}} \xrightarrow{d} N(0,1). \tag{11}$$

2.2 Considérations pratiques

Concrètement, la mise en œuvre de VSPCA suit une séquence bien définie. Tout commence par la standardisation des données témoins, suivie de l'estimation des composantes principales. Les données cas sont ensuite standardisées avec les mêmes paramètres avant d'être reconstruites. Le calcul des erreurs de reconstruction et des statistiques t_j^+ permet alors d'évaluer l'importance discriminante de chaque variable. Enfin, l'estimation robuste de la variabilité et le calcul des p-values complètent le processus.

Pour ce qui est des aspects pratiques, plusieurs considérations s'imposent. L'ajustement des p-values par la méthode FDR contrôle efficacement le taux de faux positifs. Par ailleurs, les outils diagnostiques comme les qq-plots et histogrammes des p-values offrent une validation visuelle des résultats. Notons que pour les petits jeux de données, une simple sélection basée sur le rang des statistiques peut s'avérer suffisante.

En termes de comparaison méthodologique, notre approche se mesure à trois méthodes de référence : le test t modéré (MTT) pour sa robustesse, l'analyse SAM pour son caractère non paramétrique, et la régression Lasso pour sa popularité dans la sélection de variables. Ces comparaisons permettent de situer précisément les apports et avantages de VSPCA dans le paysage méthodologique actuel.

3 Résultats

Les principaux résultats de notre méthode VSPCA, initialement présentée dans Insha et al. (2020), seront exposés en détail lors de la présentation orale. L'implémentation complète est disponible sur https://github.com/iullah1980/VSPCA. Notre approche démontre des performances remarquables pour la sélection de variables en grande dimension, surpassant les méthodes standards comme le test t modéré et l'analyse SAM sur plusieurs aspects clés.

La robustesse de VSPCA se manifeste par sa stabilité face à différents niveaux de parcimonie des données et par son contrôle efficace du taux de faux positifs. Contrairement aux méthodes de filtrage univarié qui tendent à sélectionner des variables redondantes, notre approche identifie systématiquement les marqueurs les plus discriminants tout en maintenant une excellente spécificité.

Les applications pratiques confirment ces avantages. Dans l'étude sur la leucémie aiguë lymphoblastique, VSPCA a non seulement reproduit les gènes différentiellement exprimés connus, mais a également révélé de nouvelles cibles potentielles. Son application réussie à des données de génotypage du riz démontre par ailleurs son adaptabilité à différents types de données, y compris non normales.

D'un point de vue computationnel, VSPCA présente l'avantage décisif d'une implémentation simple et légère. Alors que les méthodes comme le Lasso nécessitent des procédures de réglage intensives et des validations croisées délicates en grande dimension, notre solution s'affranchit de ces écueils tout en restant efficace même dans des situations extrêmes de déséquilibre entre variables et observations.

Bibliographie

Eckart C, Young G. The approximation of one matrix by another of lower rank. *Psychometrika* 1936; 1(3): 211–218.

Ullah, I., and Mengersen, K., and Pettitt, A., and Liquet, B., Using a supervised principal components analysis for variable selection in high-dimensional datasets reduces false discovery rates. *bioRxiv* , 2020.

Jolliffe I. *Principal component analysis.*, Springer Series in Statistics. New York: Springer Verlag. 3 ed. 2002.

Sobol-CPI : une statistique d'importance par permutation conditionnelle doublement robuste

Angel Reyero Lobo [1], Pierre Neuvial [1]& Bertrand Thirion [2]

[1] *Institut de Mathématiques de Toulouse, Université de Toulouse*
[2] *Inria, CEA, Université Paris Saclay*

Résumé. L'approche Conditional Permutation Importance (CPI) a été récemment introduite pour l'analyse de l'importance des variables, avec de bons résultats empiriques. Dans ce travail, nous fournissons d'abord des garanties théoriques pour le CPI. Nous établissons une propriété de double robustesse pour détecter les covariables nulles, ce qui en fait un modèle adapté à la sélection de variables. Nous introduisons ensuite une version modifiée et toujours efficace sur le plan computationnel, appelée *Sobol-CPI*, qui vise à estimer une mesure bien connue de l'importance des variables, l'Indice de Sobol Total (TSI). Nous prouvons qu'elle est non paramétriquement efficace et nous proposons une procédure permettant de contrôler l'erreur de type I associée. Ces résultats sont illustrés par des expériences numériques.

Mots-clés. Importance des variables, Sélection de variables, Double robustesse, Indice de Sobol total, ML interprétable.

Abstract. Conditional Permutation Importance (CPI) has been recently introduced for Variable Importance analysis with good empirical results. In this work, we first provide theoretical guarantees for CPI. We establish a double robustness property to detect null covariates, making it a suitable model for variable selection. We then present a modified and still computationally efficient version, *Sobol-CPI*, that aims to estimate a well-known variable importance measure, the Total Sobol Index (TSI). We prove that it is nonparametrically efficient, and we provide a procedure to control the type-I error. These results are illustrated by numerical experiments.

Keywords. Variable importance, Variable selection, Double robustness, Total Sobol Index, Interpretable ML.

1 Introduction

L'interprétabilité des modèles d'apprentissage automatique modernes est souvent en tension avec leur complexité. Alors que des méthodes simples comme la régression linéaire sont facilement interprétables, elles peuvent être inadaptées aux distributions sous-jacentes, produisant des résultats biaisés. Il est donc essentiel de développer des mesures de l'importance des variables qui soient *agnostiques au modèle*, capables de capturer des dépendances non linéaires et de s'adapter à différentes structures de données.

Deux grands types d'approches permettent d'évaluer l'importance d'une variable j. Les approches par *permutation* comme Permutation Feature Importance (PFI), et Conditional Permutation Importance (CPI) mesurent l'impact de la permutation de X^j sur la perte. Les approches par *suppression*, comme Leave One Covariate Out (LOCO), évaluent la performance d'un modèle ré-entraîné sans X^j.

LOCO fournit une estimation directe de l'Indice de Sobol Total (TSI) Homma and Saltelli (1996), utilisé en analyse de sensibilité. Cet indice quantifie la contribution conditionnelle d'une variable à la prédiction. Cependant, les méthodes par suppression comme LOCO ne conviennent pas aux modèles "boîte noire" car elles nécessitent un ré-entraînement. A contrario, PFI ne permet pas d'estimer correctement l'importance des variables et ne converge pas vers le TSI Bénard et al. (2022). Enfin, à ce jour, la cible théorique du CPI n'a pas été étudiée.

Nos contributions (détaillées dans l'article Reyero Lobo et al. (2025)) sont les suivantes: une analyse théorique validant l'étape d'échantillonnage conditionnel du CPI; la démonstration d'une *double robustesse* de CPI, permettant de détecter les covariables nulles même si une seule des deux estimations nécessaires est correcte, contrairement à LOCO; un lien entre les approches par suppression et permutation via *Sobol-CPI*, une nouvelle estimation du TSI, dont on montre l'efficacité asymptotique et le contrôle de l'erreur de type I; des expériences numériques illustrant ces résultats numériques.

2 État de l'art

On dispose d'un ensemble de test $(x_i, y_i)_{i=1\dots n_{\text{test}}}$ et d'un ensemble d'entraînement, tous deux échantillonnés i.i.d. selon une loi $P_0 \in \mathcal{M}$ avec $X \in \mathbb{R}^p$ et $y \in \mathbb{R}$. On étudie l'importance de la variable X^j. Soit X^{-j} le vecteur

X privé de sa j-ème coordonnée. On note $m(X) := \mathbb{E}[y|X], m_{-j}(X^{-j}) := \mathbb{E}[y|X^{-j}]$, $\nu_{-j}(X^{-j}) := \mathbb{E}[X^j|X^{-j}]$. Leurs estimées sont respectivement notées par $\widehat{m}$, $\widehat{m}_{-j}$ et $\widehat{\nu}_{-j}$. La fonction de perte est notée $\mathcal{L}$.

Total Sobol Index (TSI). L'indice de Sobol total pour la variable j quantifie la perte due à l'exclusion de X^j. Il est défini comme :

$$\psi_{TSI}(j, P_0) := \mathbb{E}\big[\mathcal{L}(m_{-j}(X^{-j}), y)\big] - \mathbb{E}\big[\mathcal{L}(m(X), y)\big].$$

On définit les méthodes existantes d'estimation de l'importance de variables.

Leave One Covariate Out (LOCO). L'estimateur LOCO (Williamson et al., 2021) est défini par:

$$\widehat{\psi}^j_{\text{LOCO}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathcal{L}\big(\widehat{m}_{-j}(x_i^{-j}), y_i\big) - \mathcal{L}\big(\widehat{m}(x_i), y_i\big).$$

LOCO nécessite d'entraîner un modèle distinct pour chaque variable, ce qui est coûteux. De plus, cette approche présente des problèmes liés aux optimisations, car les erreurs associées ne se compensent pas entre elles.

Permutation Feature Importance (PFI). L'importance par permutation (Mi et al., 2021) est donnée par:

$$\widehat{\psi}^j_{\text{PFI}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathcal{L}\big(\widehat{m}(x_i^{(j)}), y_i\big) - \mathcal{L}\big(\widehat{m}(x_i), y_i\big),$$

où $x_i^{(j)}$ est obtenu en permutant X^j. PFI est rapide à calculer, mais biaisé dans le cas de covariables corrélées (Chamma et al. (2024)).

Conditional Permutation Importance (CPI). L'importance par permutation conditionnelle (Chamma et al., 2024) est définie par:

$$\widehat{\psi}^j_{\text{CPI}} = \frac{1}{n\text{test}} \sum_{i=1}^{n_{\text{test}}} \mathcal{L}\big(\widehat{m}(\widetilde{x}_i'^{(j)}), y_i\big) - \mathcal{L}\big(\widehat{m}(x_i), y_i\big),$$

où $\widetilde{x}_i'^{(j)}$ est obtenu en permutant conditionnellement X^j. CPI corrige certains biais de PFI, mais nécessite l'estimation de ν_{-j} pour la permutation conditionnelle. Dans la suite, nous proposons une justification théorique de cette approche, ainsi qu'une correction améliorant son interprétation et sa validité statistique.

3 Double robustesse de CPI

L'estimation de deux régresseurs est nécessaire pour LOCO ($\widehat{m}$ et $\widehat{m}_{-j}$) et CPI ($\widehat{m}$ et $\widehat{\nu}_{-j}$) — pour CPI, tout échantillonneur conditionnel $\widetilde{X}'^{(j)}$ peut être utilisé à la place de $\widehat{\nu}_{-j}$. Le théorème suivant établit une propriété de double robustesse pour la détection des covariables conditionnellement nulles: contrairement à LOCO, la consistance de l'un des deux estimateurs suffit à garantir la convergence de CPI.

Theorem 3.1. *Supposons que $X^j \perp\!\!\!\perp y \mid X^{-j}$. Si l'échantillonneur conditionnel satisfait $\widetilde{X}'^{(j)} \xrightarrow{\mathcal{L}} \widetilde{X}^{(j)}$, alors, pour toute fonction de perte $\mathcal{L}$ bornée et continue et tout estimateur continu $\widehat{m}$ de m, on a $\widehat{\psi}_{\mathrm{CPI}} \to 0$. Inversement, si $\widehat{m}$ est un estimateur consistant de m, alors, pour tout échantillonneur conditionnel $\widetilde{X}'^{(j)}$ et toute perte continue $\mathcal{L}$, on a également $\widehat{\psi}_{\mathrm{CPI}} \to 0$.*

Dans un cadre linéaire, cette propriété entraîne un biais de CPI qui décroît quadratiquement avec la taille de l'échantillon d'apprentissage, tandis que pour LOCO, il ne décroît que linéairement (Reyero Lobo et al. (2025)).

4 Sobol-CPI

L'approche que nous proposons pour estimer TSI exploite la loi des probabilités totales, avec une estimation par échantillonnage conditionnel.

Definition 4.1 (Sobol-CPI)**.** *Soient $n_{\mathrm{cal}} \in \mathbb{N}^*$ et $\widehat{m}$ un régresseur de y sachant X. L'indice Sobol-CPI pour la coordonnée j associé à l'ensemble de test $(X_i, y_i)_{i=1,\dots,n_{\mathrm{test}}}$ est défini par :*

$$\widehat{\psi}^j_{\mathrm{SCPI}} = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \mathcal{L}\left(\frac{1}{n_{\mathrm{cal}}} \sum_{k=1}^{n_{\mathrm{cal}}} \widehat{m}(\widetilde{x}'^{(j)}_{i,k}), y_i \right) - \mathcal{L}(\widehat{m}(x_i), y_i),$$

où $\widetilde{x}^{(j)}_{i,k}$ est la j-ème coordonnée échantillonnée conditionnellement.

Sous certaines hypothèses de régularité et une continuité de Lipschitz, on montre que la méthode Sobol-CPI est non paramétriquement efficace, ce qui permet d'obtenir des intervalles de confiance valides. Cependant, comme pour LOCO, la fonction d'influence tend vers 0 très rapidement (Williamson et al., 2023; Verdinelli and Wasserman, 2024), ce qui empêche de donner des garanties sur l'erreur de type I. Pour obtenir de telles garanties, nous proposons une correction de la variance (Reyero Lobo et al., 2025).

5 Expériences numériques

L'un des objectifs principaux du Sobol-CPI est d'estimer l'importance des variables sans avoir à réajuster un modèle complexe pour chaque covariable. La Figure 1 est obtenue dans un cadre de simulation non-linéaire et $p = 50$. Même en utilisant un SuperLearner pour estimer $\widehat{m}$ et $\widehat{m}_{-j}$, mais un modèle efficace pour $\widehat{\nu}_{-j}$, on obtient une estimation plus rapide et de meilleurs résultats avec le Sobol-CPI. Pour les covariable importante (comme X_0 en haut à gauche), Sobol-CPI et LOCO-HD donnent des résultats similaires, mais Sobol-CPI évite le réajustement des super-learners. Pour les covariables nulles (comme X_6 en haut à droite), Sobol-CPI ne montre aucun biais. En bas à gauche, il discrimine mieux les covariables importantes. Enfin, en bas à droite, sa double robustesse est visible : avec $n_{\text{cal}} = 100$, il n'attribue toujours aucune importance aux covariables nulles.

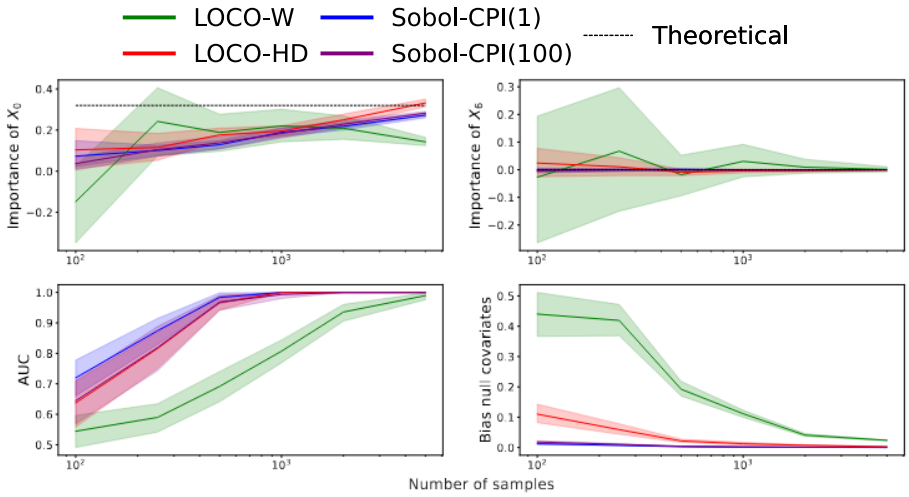


Figure 1: **Double robustesse avec des modèles complexes :** (en haut) Importance estimée : covariable significative (gauche) vs. nulle (droite). (en bas) AUC de sélection (gauche) et biais d'importance des covariables nulles (droite). Les méthodes basées sur CPI convergent généralement à des taux similaires vers TSI et, sous l'hypothèse nulle, convergent plus rapidement.

6 Conclusion

Nous avons établi une propriété de double robustesse du CPI, qui explique ses bonnes propriétés empiriques. Nous avons ensuite introduit Sobol-CPI, une modification de CPI qui cible l'Indice de Sobol Total, et prouvé son efficacité asymptotique, permettant la construction d'intervalles de confiance valides. Ces résultats sont détaillés dans Reyero Lobo et al. (2025).

References

C. Bénard, S. Da Veiga, and E. Scornet. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the sobol-mda. *Biometrika*, 109(4):881–900, 02 2022.

A. Chamma, D. A. Engemann, and B. Thirion. Statistically valid variable importance assessment through conditional permutations. *Advances in Neural Information Processing Systems*, 36, 2024.

T. Homma and A. Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17, 1996. ISSN 0951-8320.

X. Mi, B. Zou, F. Zou, and J. Hu. Permutation-based identification of important biomarkers for complex diseases via machine learning models. *Nature Communications*, 12(1):3008, 2021.

A. Reyero Lobo, P. Neuvial, and B. Thirion. Sobol-cpi: a doubly robust conditional permutation importance statistic, 2025. URL `https://arxiv.org/abs/2501.17520`.

I. Verdinelli and L. Wasserman. Feature importance: A closer look at shapley values and loco. *Statistical Science*, 39(4):623–636, 2024.

B. D. Williamson, P. B. Gilbert, M. Carone, and N. Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22, 2021.

B. D. Williamson, P. B. Gilbert, N. R. Simon, and M. Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658, 2023.

Séries temporelles

Graph-Based Weight Cascading Methods for Multisite Time Series Forecasting

Eloi Campagne[1,3] Yvenn Amara-Ouali[2,3] Yannig Goude[2,3] Mathilde Mougeot[1]
Argyris Kalogeratos[1]

[1] *Centre Borelli, Université Paris-Saclay, CNRS, Ecole Normale Supérieure Paris-Saclay*
[2] *Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, CNRS, Faculté des Sciences d'Orsay*
[3] *EDF R&D, Palaiseau*

Résumé. Cet article présente une nouvelle approche basée sur les graphes pour la prévision de séries temporelles multisites. Plus précisément, notre approche en plusieurs étapes choisit une tâche de prévision initiale, à partir de laquelle elle apprend et propage les poids appris aux tâches voisines, telles que désignées par une structure arborescente capturant les relations entre les tâches. Ce schéma de cascade des poids fonctionne en choisissant une tâche prototype comme point de départ et en affinant les poids. La méthode proposée est validée par des expériences sur des jeux de données synthétiques et réelles, démontrant une performance prédictive supérieure par rapport aux méthodes traditionnelles de raffinement, et souligne l'efficacité et l'évolutivité des méthodes basées sur les graphes dans la prévision de séries temporelles.

Mots-clés. Graphes, Apprentissage Profond, Apprentissage par Transfert, Prévision de Séries Temporelles

Abstract. This paper presents a novel graph-based approach for multisite time series forecasting. More specifically, our multistep approach chooses a starting forecasting task, from where it learns and propagates the learned weights to the neighboring tasks, as designated by a tree structure capturing the relations between tasks. This weight cascading scheme works by choosing a prototype task as a starting point and refining the weights. The proposed method is validated through experiments on synthetic and real-world datasets, demonstrating superior predictive performance compared to traditional fine-tuning methods, and highlights the efficiency and scalability and of graph-based methods in time series forecasting.

Keywords. Graphs, Deep Learning, Transfer Learning, Time Series Forecasting

[1] Mail: eloi.campagne@ens-paris-saclay.fr

1 Introduction

Transfer learning has emerged as a powerful paradigm in machine learning, allowing models to use knowledge from one task or dataset to improve performance on related tasks, especially when data is limited. In time series forecasting, this approach is particularly valuable when the historical data for each series is short, but a large number of related series are available. By capturing common patterns and dependencies across series, transfer learning improves model generalisation and forecasting accuracy. Recent advances in forecasting methods have used transfer learning and modern deep learning architectures to address these challenges. For example, PatchTST (Nie et al., 2022) segments time series into sub-series patches and employs a channel-independent design that allows for common transformer weights across univariate series. This design reduces attention map complexity, preserves local semantics, and allows for longer historical context, yielding state-of-the-art results in both supervised and self-supervised pre-training. Similarly, the MoirAI framework (Woo et al., 2024) adopts a universal prediction approach with a single pre-trained model to handle diverse tasks, addressing challenges such as cross-frequency learning and multivariate inputs. Trained on the extensive LOTSA dataset of 27 billion observations, MoirAI achieves remarkable zero-shot forecasting performance.

Beyond dealing with the above challenges of a single forecasting problem, aiming at learning jointly individual forecasting models for N time series, when N is large, is a significantly more complex and computationally expensive multitasking problem. Our approach aims to leverage the joint information of time series without resorting to a resource-intensive framework, nor to a multitasking approach that would require intertwined learning steps and iterative exchange of information among the nodes. Through this article, we seek to answer the following question: how can we achieve N accurate time series forecasts without adopting a typical multitasking approach?

2 Datasets

In this section we describe a real dataset that motivates this work, along with synthetic data that we generate to facilitate the rigorous empirical assessment of the learning approaches.

Real Dataset. The dataset[2] consists of aggregated half-hourly residential smart meter electricity consumption data collected by four UK Distribution Network Operators (DNOs). Currently, data from the two largest DNOs, Scottish and Southern Electricity Networks (SSEN) and National Grid Electricity Distribution (NGED), are integrated into an efficient GeoParquet[3] format, covering approximately 2 million smart meters, $120,000$ low voltage feeders and $50,000$ substations.

[2]https://weave.energy/
[3]https://geoparquet.org/

Synthetic Dataset. To generate a dataset with spatial correlations, we developed a time series model whose parameters vary smoothly across a geographic landscape. The model is defined as:

$$y(t) = m + s[\sin(\omega_1 - \varepsilon_1)t + \sin(\omega_7 - \varepsilon_7)t] + \mathcal{N}(0, 100), \tag{1}$$

where $y(t)$ denotes the time series indexed by time t, m is an intercept term, s acts as a scaling parameter, and ω_1, ω_7 are respectively daily and weekly pulsations. The terms $\varepsilon_1, \varepsilon_7$ represent perturbations that introduce stochasticity into the frequency components. To impose spatial coherence, we constructed four spatial fields, each controlling one of the parameters. These fields were defined over a unit square $[0, 1] \times [0, 1]$ domain using a structured meshgrid. To generate smoothly varying parameter landscapes, we approximated 4D Perlin noise (Perlin, 2002, Rose and Bakaoukas, 2016) by using two 3D Perlin noise functions, effectively projecting a higher-dimensional noise process onto a two-dimensional spatial manifold. The resulting landscapes were then rescaled to match predefined parameter ranges. Finally, we randomly sampled spatial locations within the domain, each associated with a unique parameter set, see Figures 1 and 2.

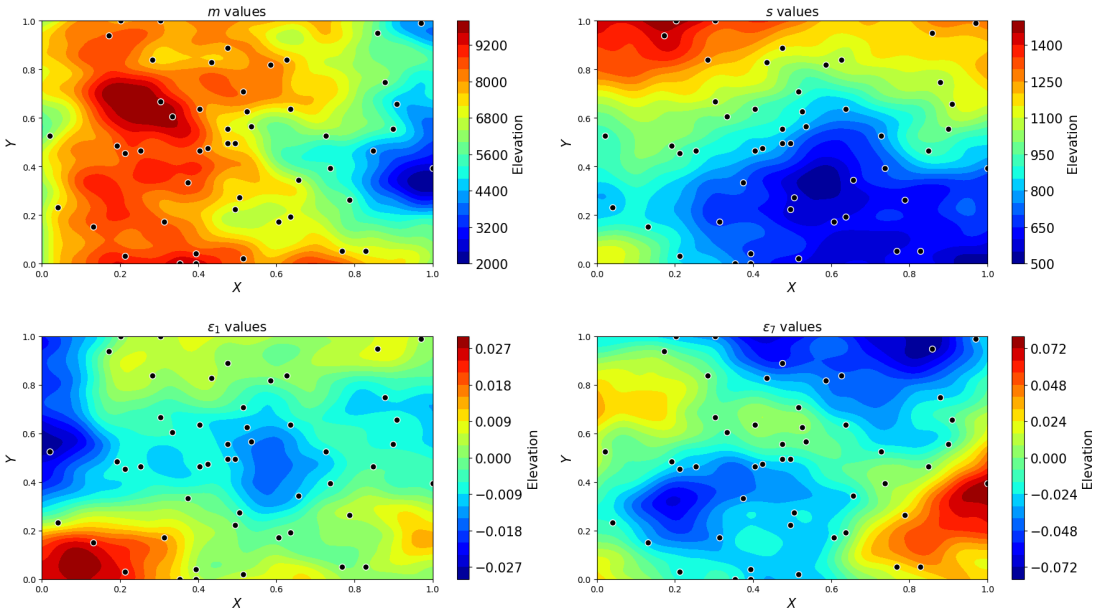


Figure 1: $N = 50$ sampled points on generated smooth univariate landscapes.

3 Contributions and Results

Let $\mathcal{D}$ be a time-indexed graphical dataset, where each $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$ is a graph defined by a fixed set of nodes $\mathcal{V}$ and edges $\mathcal{E}_t$ that vary over time. We consider the setting where the number of time indices T is relatively small, while the number of nodes $|\mathcal{V}|$ is large. This necessitates information sharing across nodes to ensure accurate task execution. Let $\mathcal{F}$ denote a set of parametric functions $f_{\boldsymbol{\theta}}$ mapping from a domain E to some range, where $\theta \in \Theta$ represents the parameter vector to be optimized. For example, in a time series forecasting

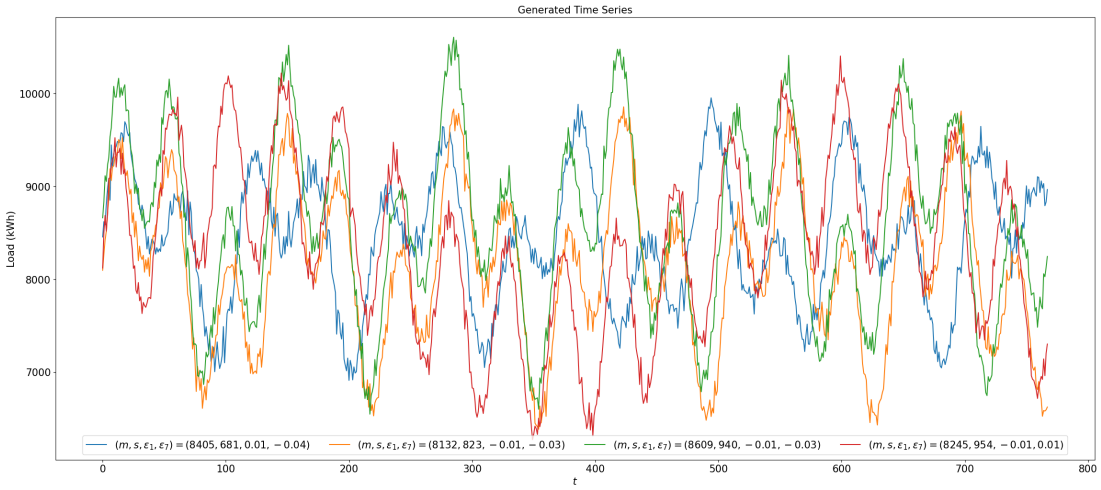


Figure 2: Four Generated time series out of the sampled points.

task, the domain and range are subsets of $\mathbb{R}^T$. To enable information sharing across nodes, we define a task-specific distance measure. For a given task $\mathfrak{T}$, we introduce a distance function $d_{\mathfrak{T}} : E \times E \to \mathbb{R}_+$. For instance, in time series forecasting, the distance $d_{\mathfrak{T}}$ can be chosen as the Euclidean distance. We want to build a graph structure that allows us to efficiently propagate information between tasks, which can be ensured by minimizing the distances $d_{\mathfrak{T}}$ on the graph.

Contributions. We introduce a new approach to fine-tuning the weights of several individual models via cascade diffusion over an optimal branching. In particular, we show that this method is of lower time complexity and outperforms other fine-tuning approaches in terms of forecast accuracy, both on the real electricity consumption dataset and on the synthetic toy dataset. Our approach uses a graph structure to transfer knowledge across time series and exploit their connections to build more robust models when there are limitations in the amount of available historical data. The resulting graph-based framework is compelling for knowledge transfer across several time series, with either Minimum Spanning Trees (MSTs) or Minimum Cost Arborescences (MCAs) being natural choices to drive the process depending on the nature of the task dependencies. For undirected graphs, relationships are symmetric and MSTs are particularly advantageous, as they optimize cost efficiency while ensuring full connectivity. As an interesting analogue to our learning strategy and the use of a tree structure to drive the process, we can mention that MSTs have been effectively used in engineering contexts such as the one of urban water supply networks where the cost of pipeline installation needs to be minimized while making sure that all sites of a region are connected to the network (Cong and Zhao, 2015). By considering a distance graph between the different sites and then applying algorithms such as Kruskal's, a MST can be constructed to ensure the lowest cost while meeting safety, reliability, and economic requirements. In the context if time series forecasting, when the dependencies between the time series are asymmetric, MCAs provide an adapted framework by ensuring that every node is reachable from a designated root while minimizing the total edge cost.

Results. Our initial experiments reveal several insights: first, by refining models using our algorithms that exploit the underlying graph structure of the data, we consistently achieve performance similar to individually trained models while significantly reducing computation time. Furthermore, when we infer an optimal graph structure for weight refinement, the performance not only surpasses that of individual models but also outperforms traditional refinement methods.

4 Perspectives

Our algorithms are currently in the adjustment phase, particularly for inferring a graph suitable for refining model weights. The tree graphs that we compute (MSTs or MCAs) perform significantly better than random tree graphs, which substantiates the idea of finding a proper graph to drive the multistep process of learning and propagating weights among tasks. However, w observe that the process is sensitive to the tree graph used, meaning that small variations in an otherwise well-performing tree can lead to a deterioration in results.

References

Cong, F. and Zhao, Y. (2015). The application of minimum spanning tree algorithm in the water supply network. In *International Industrial Informatics and Computer Engineering Conference*. Atlantis Press.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. (2022). A time series is worth 64 words: Long-term forecasting with transformers.

Perlin, K. (2002). Improving noise. In *Proceedings of the Conference on Computer graphics and Interactive Techniques*.

Rose, T. J. and Bakaoukas, A. G. (2016). Algorithms and approaches for procedural terrain generation-a brief review of current techniques. In *IEEE International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. (2024). Unified training of universal time series forecasting transformers.

Change Point Detection in Hadamard Spaces by Alternating Minimization

Anica Kostic [1], Vincent Runge [2] & Charles Truong [3]

[1] *Department of Statistics, London School of Economics and Political Science, a.kostic@lse.ac.uk*
[2] *Laboratoire de Mathématiques et Modélisation d'Evry, Université Paris-Saclay, CNRS, Univ Evry, vincent.runge@univ-evry.fr*
[3] *Centre Borelli, Université Paris-Saclay, CNRS, ENS Paris-Saclay, ctruong@ens-paris-saclay.fr*

Abstract. Time series analysis of non-Euclidean data is highly challenging and crucial for many real-world applications. We address the problem of detecting multiple changes in time series within these complex data spaces. Hadamard spaces, which encompass important data spaces like positive semidefinite matrices, certain Wasserstein spaces, and hyperbolic spaces, provide the right general framework to address this complexity. We propose a computationally efficient two-step iterative optimization algorithm called HOP (Hadamard Optimal Partitioning) that detects changes in the sequence of so-called Fréchet means. Under mild conditions, the proposed method consistently estimates the change point locations. HOP is highly versatile, accommodating structural assumptions such as cyclic patterns and epidemic settings, making it unique in the literature. We validate its performance in synthetic and real-world scenarios, including applications in human gait analysis using EMG data with low SNR and behavioral analysis of animal motion.

Keywords. Change point detection, Hadamard space, Frechet mean, Dynamic programming, Time series analysis

1 Introduction and Problem Statement

Detecting change points in time series is a longstanding statistical challenge. Traditional approaches rely on parametric models from the exponential family, offering strong theoretical guarantees and efficient algorithms. Recent interest has shifted towards change point detection in non-Euclidean spaces, motivated by applications such as: 1) Dynamic covariance matrices in neuroscience; 2) Time series of probability distributions, e.g., stock return distributions; 3) Dynamic graphs, modeling evolving networks. In these settings, data reside in metric spaces with non-Euclidean geometry, requiring new methodologies. While statistical tools for such data are developing, methods for change point segmentation of such data are very few (Dubey and Müller, 2020; Jiang et al., 2024).

We address the multiple change point problem in general metric spaces, focusing on Hadamard spaces, where the Fréchet mean is well-defined and unique (Sturm, 2002). We propose HOP, a computationally efficient two-step optimisation algorithm with theoretical

guarantees for change point detection and localisation. The method is validated on both synthetic and real data, including cyclic EMG signals, where it successfully identifies meaningful patterns.

Background on Hadamard Spaces A metric space $(\mathcal{H}, d)$ is called a Hadamard space if it is complete and satisfies for each pair of points $(x_0, x_1) \in \mathcal{H}^2$, there exists a point $m \in \mathcal{H}$ such that for all $y \in \mathcal{H}$:

$$d^2(m, y) + \frac{1}{4}d^2(x_0, x_1) \leq \frac{1}{2}d^2(y, x_0) + \frac{1}{2}d^2(y, x_1). \tag{1}$$

These spaces, also called global NPC (Non-Positive Curvature) spaces, have wide applicability in various fields. We consider applications in spaces such as the space $\mathcal{S}_p$ of symmetric positive definite matrices, with the affine invariant metric, and metric trees, equipped with their geodesic distance.

Without the familiar Euclidean structures, basic concepts like the mean must be redefined, and existence or uniqueness isn't always assured. The Fréchet mean of a set of points $y_1, \ldots, y_T$ in a Hadamard space is defined as $\hat{\mu} = \arg\min_\mu \sum d^2(y_i, \mu)$. The favorable regularity properties of Hadamard spaces guarantee the existence of a unique mean. As a result, calculating the Fréchet mean becomes a convex optimization problem.

Model and Problem Statement Consider a Hadamard space $(\mathcal{H}, d)$ and a $\mathcal{H}$-valued signal $\boldsymbol{y} = (y_t)_{t=1}^T$ generated by a random vector $(Y_t)_{t=1}^T$, where the expectation $\theta_t^* := \mathbb{E}Y_t$ is piecewise constant. We define the change points as indices $t_1^\star, \ldots, t_{K^\star}^\star$ such that $\theta_t^\star = c_k^\star$ for $t_k < t \leq t_{k+1}^\star$.

$$\theta_t^\star = c_k^\star \quad \text{if} \quad t_k^\star < t \leq t_{k+1}^\star\,,\ k = 0, \ldots, K^\star\,, \tag{2}$$

where $c_k^\star \in \mathcal{H}$ $(k = 0, \ldots, K^\star)$ are the successive Fréchet means and, $t_0^\star := 0$ and $t_{K^\star+1}^\star := T$. The number of different mean values is $\#\{c_k^*,\, k = 0, \ldots, K^\star\} = \upsilon \leq K^\star + 1$. We define $\hat{\boldsymbol{\theta}} = (\hat{\theta}_t)_{t=1}^T$, as the estimator of $(\boldsymbol{\theta}^\star)_{t=1}^T$, which is obtained by minimizing a penalized criterion based on an L_0 penalty:

$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta} \in \mathcal{H}^T}{\arg\min} \left[\sum_{t=1}^T d^2(y_t, \theta_t) + \beta \sum_{t=2}^T \mathbf{1}_{\{\theta_{t-1} \neq \theta_t\}} \right], \tag{3}$$

where $\mathbf{1}$ is the indicator function and $\beta > 0$ controls the trade-off between data fidelity and the number of changes. Other penalty functions, not proportional to the number of changes, can also be used.

This estimator is a piecewise constant sequence of states, decomposing the signal into segments. On an interval between two change-points, a constant representative value c is found to minimise the sum of squared distances, $\min_{y \in \mathcal{H}} \left\{ \sum_{t=i}^j d^2(y_t, y) \right\}$. The solution is unique due to the Hadamard structure. The dynamic programming algorithm can be used to solve the minimization problem (3). It achieves quadratic time complexity assuming that segment costs can be computed in constant time, as in exponential family models. However, computing the Fréchet mean in Hadamard spaces is more challenging. Polynomial-time algorithms exist for specific spaces, but these can still be time-consuming for large datasets.

2 The HOP Method

To efficiently approximate the solution to problem (3), we propose an alternating algorithm, HOP (Hadamard Optimal Partitioning). We first choose the number of states, V, and the initial set of values, denoted as $\boldsymbol{c}^{(0)}$, (typically using an informed approach, such as k-means). This known set of states is then used to compute an initial path estimator, $\boldsymbol{p}^{(1)} \in \{1, \dots, V\}^T$, which is subsequently used to refine the set of states, and so on. The estimator is progressively improved with each iteration of the (fast) dynamic programming algorithm. We define:

$$\Phi : \begin{cases} \mathcal{H}^V \times \{1, \dots, V\}^T \longrightarrow \mathbb{R} \\ (\boldsymbol{c}, \boldsymbol{p}) \longmapsto \displaystyle\sum_{t=1}^{T} d^2(y_t, c_{p(t)}) + \beta \sum_{t=2}^{T} \mathbf{1}_{\{p(t-1) \neq p(t)\}} \,. \end{cases} \tag{4}$$

The sequence of alternating problems for inferring the path and the state values is defined recursively for $i = 1, 2 \dots$ as:

$$\begin{cases} \boldsymbol{p}^{(i+1)} := \displaystyle\operatorname*{arg\,min}_{\boldsymbol{p} \in \{1, \dots, V\}^T} \Phi(\boldsymbol{c}^{(i)}, \boldsymbol{p}) \,, \\ \boldsymbol{c}^{(i+1)} := \displaystyle\operatorname*{arg\,min}_{\boldsymbol{c} \in \mathcal{H}^V} \Phi(\boldsymbol{c}, \boldsymbol{p}^{(i+1)}) \,. \end{cases} \tag{5}$$

The update $\boldsymbol{p}^{(i+1)}$ is found using the dynamic programming algorithm. The update $\boldsymbol{c}^{(i+1)}$ consists in finding the Fréchet mean state by state, that is, for the j-th state, by solving:

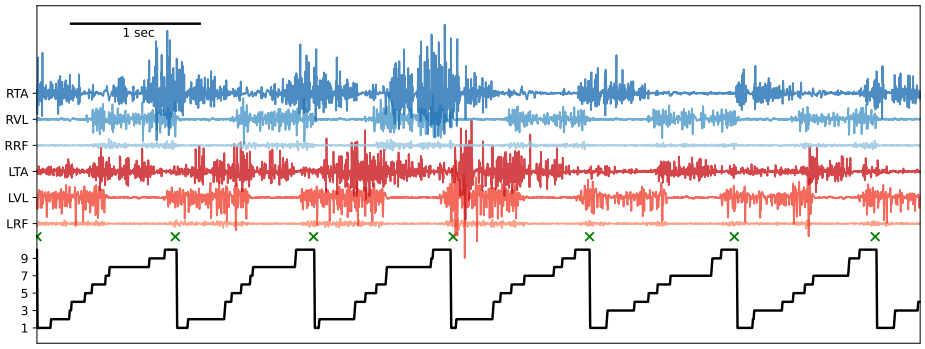
$$c_j^{(i+1)} = \operatorname*{arg\,min}_{\mu \in \mathcal{H}} \sum_{t=1}^{T} \mathbf{1}_{\{p_t^{(i+1)} = j\}} d^2(\mu, y_t).$$

If the sequences $\boldsymbol{p}^{(i)}$ and $\boldsymbol{p}^{(i+1)}$ are identical, the algorithm has converged to a (local) minimizer, which defines our final estimator $\left(c_{p_t^{(i)}}^{(i)}\right)_{t=1}^{T}$.

3 Applications

3.1 Human Gait Analysis With EMG Signals

Our aim in this application is to decode long, noisy EMG time series collected from subjects walking in different manners (e.g., backward walk, heel walking, toe walking). Starting from the raw signal data collected from 16 sensors, we compute a sliding covariance matrix, yielding $\mathcal{S}_{16}$-valued signals containing around 4000 observations each. The HOP algorithm is applied to this data. To facilitate the analysis, we enforce a cyclic structure of state transitions through the penalty function of our method, reflecting the cyclic nature of human locomotion. The results, shown in Figure 1, demonstrate that HOP achieves a 94.7% F-score, effectively recovering footsteps across walking styles.



(a)

Precision	$94.2 \pm 1.5\%$
Recall	$95.2 \pm 1.3\%$
F-score	$94.7 \pm 1.4\%$

(b)

Figure 1: (a) EMG signal example (8 seconds). Only 6 dimensions out of 16 are shown from sensors placed on the tibialis anterior (TA), vastus lateralis (VL), and the rectus femoris (RF) on the left leg ("L", in red), and the right leg ("R", in blue). Here, the subject walks backward; the green crosses delimit the true steps. The black curve shows the sequence of states learned by our method. The true steps coincide with the change from State 10 to State 1. (b) Performance of HOP on the EMG data set (margin 75 ms).

3.2 Behavioral Analysis of Animal Motion

In this section, we use HOP to analyze rodent motion video data. The information of the original videos is compressed into a sequence of so-called action syllables known as ethograms. These syllables include actions like "turning to the right", "standing still", "rearing". The ethograms can be further organised into a hierarchical structure, a tree, where the leaves are action syllables, while the parents, called meta-syllables, represent more complex actions, such as stationery, grooming, exploration. This tree structure is an example of a Hadamard space. We apply HOP on the ethogram for different penalty values, with higher penalties correspond to more filtered ethograms. The results are shown in Figure 2. In summary, we compress the original ethogram, which contains 50 distinct syllables (Figure 2-b), into a high-level ethogram with only 7 distinct meta-syllables (Figure 2-c). This filtered ethogram is easier to understand in one glimpse. We highlight two of the meta-syllables found and show that they are interpretable, as they are related to two well-known behaviors: exploration and grooming.

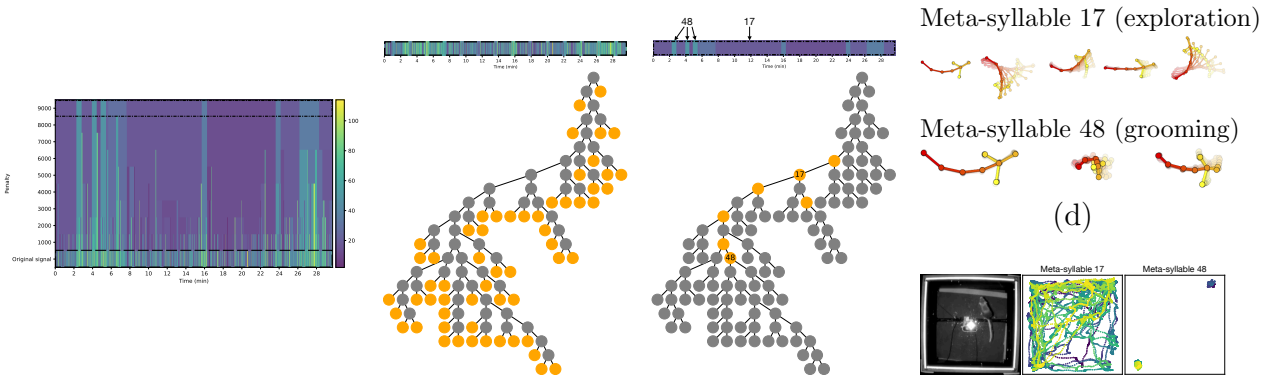


Figure 2: (a) Filtered ethograms for several penalty values. The original (unfiltered) ethogram is at the bottom; the most filtered is at the top. Each color is an action syllable. (b) The original ethogram is shown, and below is the hierarchical representation of action syllables. Orange nodes correspond to syllables that are present in the original ethogram. By construction, they correspond to leaves. (c) The most filtered ethogram is shown, and below, the hierarchical representation of action syllables. Orange nodes correspond to syllables that are present in the most filtered ethogram. The orange nodes are closer to the root node, meaning that they encode higher-level semantics. Two particular (meta-)syllables, 17 and 48, are highlighted in the ethogram and the tree. (d) Most frequent action syllables associated with Meta-syllables 17 and 48, which respectively correspond to "exploration" and 'grooming". For each action syllable, the average 8-point mouse skeleton is shown. (e) Image of the mouse and its cage, and scatter plot of mouse's center of mass while in Meta-syllables 17 and 48. Trajectories here are color-coded: the starting (resp. end) point is depicted in blue (resp. yellow). This representation confirms that Meta-syllable 17 corresponds to exploration (the mouse went over the whole cage) and Meta-syllable 48 to grooming (the mouse stayed in the corners).

References

Paromita Dubey and Hans-Georg Müller. Fréchet change-point detection. *The Annals of Statistics*, 48(6):3312–3335, 2020.

Feiyu Jiang, Changbo Zhu, and Xiaofeng Shao. Two-sample and change-point inference for non-euclidean valued time series. *Electronic Journal of Statistics*, 18(1):848–894, 2024.

Karl-Theodor Sturm. Probability measures on metric spaces of nonpositive curvature. In *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces*, volume 338, pages 357–390. American Mathematical Society, Paris, France, contempora edition, 2002.

Estimation robuste en ligne dans le cas du modèle espace-état linéaire avec covariables stochastiques

Jean-Luc MAHOROMEZA [1,2] & Olivier WINTENBERGER [1] & Adeline FERMANIAN [2] & Joseph DE VILMAREST [3]

[1] *Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Paris, France*
[2] *LOPF, Califrais' Machine Learning Lab, Paris, France*
[3] *Viking Conseil, Paris, France*

Résumé. Le filtre de Kalman-Bucy permet l'estimation et la prédiction des modèles espace-état. Il est souvent confronté à des observations incertaines, par exemple quand elles proviennent de mesures bruitées d'un signal physique (capteurs, météo, etc.). Plutôt que de considérer les covariables comme des quantités déterministes, ce qui est fait classiquement, notre objectif est de prendre en compte leurs incertitudes et d'étudier la robustesse du filtre. Plus précisément, le filtre de Kalman est une approche bayésienne qui donne l'estimation de la loi *a posteriori* de l'état, conditionnellement aux observations passées. Nous raffinons cette approche en intégrant les incertitude des covariables dans le modèle, c'est-à-dire en les considérant comme des variables aléatoires dont la loi doit être estimée.

Mots-clés. Robustesse, Séries temporelles, Filtre de Kalman, Statistiques bayésiennes.

Abstract. The Kalman-Bucy filter is used to estimate and predict state-space models. It frequently faces uncertain observations, such as those derived from noisy measurements of physical signals (including sensors, weather data, etc.). Instead of considering covariates as non-stochastic quantities, as is conventionally done, our purpose is to take their uncertainties into account and address a robust filter. More specifically, the Kalman filter is a Bayesian approach that estimates the posterior distribution of the state conditionally on past observations. We refine this approach by integrating the uncertainties of the covariates into the model, i.e., considering them as stochastic variables whose distributions need to be estimated.

Keywords. Robustness, Time series, Bayesian Filtering, Bayesian statistics.

1 Introduction

Les modèles espace-état se sont imposés comme un cadre incontournable pour analyser les modèles dynamiques. Ces modèles reposent sur l'idée qu'un vecteur observable y_t, souvent appelé "vecteur d'observation", est une fonction du vecteur non observé (latent) θ_t, désigné comme le "vecteur d'état". La dynamique régissant le vecteur d'observation et le vecteur d'état est encapsulée au sein du modèle espace-état. Dans la seconde moitié du 20ème siècle, le filtre de Kalman and Bucy (1961) a considérablement avancé la modélisation des modèles

espace-état. Il a fourni un cadre d'estimation des modèles linéaires, fondé sur le filtre de Wiener (Wiener, 1949). Le filtre de Kalman-Bucy se distingue par son mécanisme récursif pour estimer et prédire l'état des modèles dynamiques linéaires dans un contexte de données bruitées. Cet avancement a été crucial pour la prédiction précise (l'espérance conditionnelle de l'état sachant le passé est calculée de façon exacte), l'estimation et le contrôle dans les modèles dynamiques linéaires, marquant une réalisation significative dans l'ingénierie, y compris le suivi de cibles, le contrôle d'avions (Ray and Stengel, 1991), l'ingénierie électrique (Vilmarest, 2022) et les sciences environnementales.

Au fil du temps, le filtre de Kalman a connu un succès généralisé, menant à diverses interprétations. Ollivier (2018) a établi un lien entre le filtre de Kalman étendu (Fahrmeir, 1992) et le gradient naturel en ligne de Murata and Amari (1999). De plus, Durbin and Koopman (2012) définit le problème de filtrage dans une configuration gaussienne linéaire comme un problème de recherche d'un estimateur linéaire non biaisé de variance minimale. Plusieurs variantes ont été développées pour adapter le filtre de Kalman à une gamme de problèmes. Ces adaptations ont exploré de nombreuses directions, incluant : (1) l'extension du filtrage aux scénarios non linéaires (Wan and Van Der Merwe, 2000; Fahrmeir, 1992); (2) le raffinement du filtre pour les cas où les hyper-paramètres du modèle sont inconnus (de Vilmarest and Wintenberger, 2021); (3) l'amélioration de la robustesse du filtre face aux perturbations du modèle, incluant la prise en compte à la fois du bruit statistiquement inconnu au sein du modèle espace-état et des incertitudes dans les variables du modèle dues à des erreurs déterministes ou stochastiques (Yedavalli, 2014; Zeng et al., 2012; Ra and Whang, 2008; Theodor and Shaked, 1996; Stengel and Ray, 1991; Ray and Stengel, 1991); et (4) la prise en compte des erreurs de modélisation où le modèle idéal est souvent mal défini, nécessitant parfois des hypothèses trop fortes qui peuvent s'écarter nettement des conditions réelles (Toda and Patel, 1980; Nishimura, 1970).

Notre étude aborde le problème (3), se concentrant sur la construction d'un estimateur robuste aux perturbations qui se manifestent comme du bruit dans les covariables. De même, notre étude s'inscrit dans la direction du problème (4) ; l'hypothèse courante que les covariables sont déterministes néglige leur nature intrinsèquement bruitée ou estimée. L'étude de la robustesse d'un estimateur se divise en deux grandes catégories : la première suppose une perturbation déterministe, sans aucun modèle probabiliste attribué à cette perturbation (Yedavalli, 2014). La seconde catégorie couvre les systèmes soumis à des perturbations stochastiques (Ra and Whang, 2008; Theodor and Shaked, 1996; Stengel and Ray, 1991).

Dans notre étude, nous considérons un modèle bien spécifié mais dans lequel les covariables sont bruitées. Ce cas des covariables soumises aux perturbations stochastiques se rencontre dans un large spectre d'applications. Par exemple, dans les sciences environnementales, elle est essentielle pour modéliser la dispersion des polluants et les prévisions du changement climatique, où les covariables comme les concentrations de polluants ou les mesures de température viennent avec leur propre ensemble d'incertitudes. Le secteur financier s'appuie sur des méthodologies similaires pour modéliser les indicateurs économiques et les prix des actifs, en intégrant la volatilité et les évaluations des risques pour prendre des décisions d'investissement éclairées. Dans le domaine de la santé, en particulier dans le domaine de l'épidémiologie, la modélisation de la propagation des maladies avec des données

incertaines est critique pour les stratégies de planification et de réponse. De plus, l'ingénierie, en particulier les systèmes de contrôle et la robotique, applique ces principes pour compenser les imprécisions des capteurs et pour affiner les réponses du système. La météorologie se démarque également, où la précision des prédictions météorologiques est considérablement améliorée en modélisant la perturbation dans les données atmosphériques.

2 Setting and preliminaries

We consider a univariate time series forecasting problem. We denote by $y_t \in \mathbb{R}$ the variable of interest at time t. In this paper, we set aside the standard linear Gaussian state-space model where the observation y_t depends linearly on a parameter θ_t and deterministic vector $x_t \in \mathbb{R}^d$, commonly called covariates and move on to a more generalized case by allowing for the possibility that **the covariates are stochastic**.

More specifically, given θ_0, we consider a Gaussian linear time-varying state-space model $\forall t \geq 1$:

$$\begin{aligned} &\text{State:} && \theta_t = A\theta_{t-1} + \eta_t, \\ &\text{Observation:} && y_t = x_t^\top \theta_t + \varepsilon_t, \\ &\text{Covariate prior:} && x_t = \hat{x}_t + \xi_t. \end{aligned} \tag{1}$$

The hidden state $\theta_t \in \mathbb{R}^d$ is to estimate, and $y_t \in \mathbb{R}$ is the observation to forecast. The transition from θ_{t-1} to θ_t is governed by a known matrix $A \in \mathbb{R}^{d\times d}$. The noises ε_t and η_t are mean-zero independent Gaussian processes with known variances σ_t^2 and Q_t, respectively. The noise ξ_t is a mean-zero independent process with known variance Ω_t but is not necessary Gaussian. In our framework, we possess for each time t, an unbiased estimator $\hat{x}_t$ of the covariates vector x_t.

In this paper, we make the following assumptions:

Assumption 1 *The initial condition $(\hat{\theta}_0, P_0)$ is known.*

Assumption 2 *For all $t \geq 1$, η_t, ξ_t, and ε_t are mutually independent.*

Let $v_{1:t}$ represent the sequence $\{v_1, \ldots, v_t\}$. For any s and t such that $t \leq s$, we define

$$\hat{\theta}_{t|s} = \mathbb{E}\left[\theta_t \mid y_{1:s}\right], \tag{2a}$$

$$P_{t|s} = \mathbb{E}\left[(\theta_t - \hat{\theta}_{t|s})(\theta_t - \hat{\theta}_{t|s})^\top \mid y_{1:s}\right], \tag{2b}$$

$$\tag{2c}$$

and observation estimator and variance by

$$\hat{y}_{t|s} = \mathbb{E}[y_t \mid y_{1:s}], \tag{3a}$$

$$\sigma^2_{t|s} = \mathbb{E}\left[(y_t - \hat{y}_{t|s})^2 \mid y_{1:s}\right]. \tag{3b}$$

We aim to design and assess estimators that are "robust" with respect to the variance Ω_t of the covariates. In line with the robustness perspective of Huber and Ronchetti (2009), robustness focus on three key properties:

- Efficiency: The estimator should be near-optimal (or reasonably good) under the deterministic case ($\Omega_t = 0$).
- Stability: Small perturbations or "uncertainty", Ω_t close to "zero", should have only a minor negative effect on the estimator's performance.
- Breakdown: Large perturbations or "uncertainty", $\Omega_t \to \infty$, should not lead to "catastrophic" failure or unbounded errors.

When covariates are deterministic, the posterior distribution of θ_t is Gaussian. Apart from the fact that the Kalman Filter is extremely sensitive to deviations from the linear Gaussian model and therefore not robust, when covariates are stochastic, the posterior state distribution is not Gaussian any more and the Kalman Filter is not optimal.

3 Theoretical results

3.1 Space prediction

Theorem 3.1 *Consider the state-space model* (1)*. The following results hold for all* $t > 0$*:*

$$\begin{aligned}\hat{y}_{t|t-1} &= \hat{x}_t^\top \hat{\theta}_{t|t-1},\\ \sigma^2_{t|t-1} &= \hat{x}_t^\top P_{t|t-1}\hat{x}_t + \operatorname{Tr}\left(\Omega_t V_{t|t-1}\right) + \sigma_t^2,\end{aligned}$$

where $V_{t|t-1} = P_{t|t-1} + \hat{\theta}_{t|t-1}\hat{\theta}^\top_{t|t-1}$.

Theorem 3.1 shows that the mean and the variance estimators are the same as the Kalman Filter estimators with an irreducible extra term $\operatorname{Tr}\left(\Omega_t V_{t|t-1}\right)$ from the covariates and states variances in $\sigma^2_{t|t-1}$.

3.2 State estimation

3.2.1 Linear estimation (MMSE)

Consider estimators of the form:

$$\mathcal{C}_t^{\text{lin}} = \left\{\beta_t + \gamma_t\, y_t \mid \beta_t, \gamma_t \in \mathbb{R}^d\right\}, \tag{4}$$

where β_t and γ_t are sought such that they represent the coordinates of the orthogonal projection of θ_t onto the span of $\{1, y_t\}$ in the space $L^2\big(\mathbb{R}^d, \sigma(y_{1:t-1}), \mu\big)$. Here, μ is a probability measure absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$.

Let $\hat{\theta}_t \in \mathcal{C}_t^{\text{lin}}$ be a square-integrable estimator of θ_t. We define the function $\mathbf{R}_{\theta_t}$ of $\mathcal{C}_t^{\text{lin}}$ in $\mathcal{S}_+^d$, where $\mathcal{S}_+^d$ is the space of symmetric positive semidefinite matrices, as:

$$\mathbf{R}_{\theta_t}(\hat{\theta}_t) = \mathbb{E}\left[(\theta_t - \hat{\theta}_t)(\theta_t - \hat{\theta}_t)^\top \middle| y_{1:t-1}, \theta_t\right]. \tag{5}$$

Using the bias-variance decomposition, $\mathbf{R}_{\theta_t}$ can be rewritten as

$$\mathbf{R}_{\theta_t}(\hat{\theta}_t) = \mathbb{V}(\hat{\theta}_t) + \mathbf{b}(\hat{\theta}_t)\mathbf{b}(\hat{\theta}_t)^\top, \tag{6}$$

where the bias $\mathbf{b}(\hat{\theta}_t)$ and the variance $\mathbb{V}(\hat{\theta}_t)$ are given by:

$$\mathbf{b}(\hat{\theta}_t) = \theta_t - \mathbb{E}\left[\hat{\theta}_t \mid y_{1:t-1}, \theta_t\right], \tag{7a}$$

$$\mathbb{V}(\hat{\theta}_t) = \gamma_t \, \mathbb{E}\left[\left(y_t - \hat{x}_t^\top \theta_t\right)^2 \middle| y_{1:t-1}, \theta_t\right] \gamma_t^\top. \tag{7b}$$

The function $\mathbf{R}_{\theta_t}$ characterizes the shape, orientation, and scale of the uncertainty associated with $\hat{\theta}_t$, it quantifies in a certain way the risk of the estimator $\hat{\theta}_t$. Geometrically, this dispersion can be visualized as an ellipsoid centred at θ_t, where the directions correspond to eigenvectors and the lengths are proportional to eigenvalues. Thus, a concentrated ellipsoid with small eigenvalues across all directions, can be considered to lower risk in the estimator $\hat{\theta}_t$ within the set $\mathcal{C}_t^{\text{lin}}$.

If we consider that $\hat{\theta}_t \in \mathcal{C}_t^{\text{lin}}$ as a pair $\hat{\theta}_t = (\beta_t, \gamma_t)$ where β_t, γ_t are $\sigma\{y_{1:t-1}\}$-measurable, we denote $\forall t \geq 1$,

$$\mathbf{R}(\beta_t, \gamma_t) = \mathbb{E}\left[\mathbf{R}_{\theta_t}(\beta_t + \gamma_t y_t) \mid y_{1:t-1}\right].$$

Theorem 3.2 *Consider the state-space model* (1). *We have* $\forall \beta_t, \gamma_t \in \mathbb{R}^d$ *and* $\sigma\{y_{1:t-1}\}$-*measurable,*

$$\mathbf{R}(\beta_t, \gamma_t) - \mathbf{R}(\beta_t^{\text{MV}}, \gamma_t^{\text{MV}}) \succeq 0,$$

if and only if

$$\beta_t^{\text{MV}} = (Id - \gamma_t^{\text{MV}} \hat{x}_t^\top)\hat{\theta}_{t|t-1},$$
$$\gamma_t^{\text{MV}} = \frac{P_{t|t-1}\hat{x}_t}{\hat{x}_t^\top P_{t|t-1}\hat{x}_t + \text{Tr}\left(\Omega_t V_{t|t-1}\right) + \sigma_t^2}.$$

with

$$\mathbf{R}(\beta_t^{\text{MV}}, \gamma_t^{\text{MV}}) = \left(I_d - \gamma_t^{\text{MV}} \, \hat{x}_t^\top\right) P_{t|t-1}.$$

The vector β_t^{MV} and γ_t^{MV} in Theorem 3.2, define an estimator $\hat{\theta}_{t|t}^{\mathrm{MV}} \in \mathcal{C}_t^{\mathrm{lin}}$ which is the best linear trade-off bias-variance. Since $\hat{\theta}_{t|t}^{\mathrm{MV}}$ achieves a lower risk $v^\top \mathbb{E}\big[\mathbf{R}_{\theta_t}\big]\, v$ with respect to any direction $v \in \mathbb{R}^d$, it minimizes the mean square error in particular. Thus we define the minimum mean square error $MMSE$ estimator and its risk as

$$\hat{\theta}_{t|t}^{\mathrm{MMSE}} = \hat{\theta}_{t|t-1} + \gamma_t^{\mathrm{MMSE}}\, y_t, \tag{8a}$$

$$\mathbf{R}_{t|t}^{MMSE} = \big(I_d - \gamma_t^{\mathrm{MMSE}}\, \hat{x}_t^\top\big)\, P_{t|t-1}, \tag{8b}$$

where

$$\gamma_t^{\mathrm{MMSE}} = \frac{P_{t|t-1}\hat{x}_t}{\hat{x}_t^\top P_{t|t-1}\hat{x}_t + \mathrm{Tr}\left(\Omega_t V_{t|t-1}\right) + \sigma_t^2}.$$

The estimator $\hat{\theta}_{t|t}^{\mathrm{MMSE}}$ and its variance $\mathbf{R}_{t|t}^{MMSE}$ closely resemble the form of the Kalman Filter $\hat{\theta}_{t|t}^{\star}$ and $P_{t|t}^{\star}$. The main differences are the use of $\hat{x}_t$ in place of x_t, and the adjustment of the gain by $\mathrm{Tr}\left(\Omega_t V_{t|t-1}\right)$, which accounts for the stochastic contribution of the covariates. This adjustment is intuitive: when the variance Ω_t of the covariates increases, the gain γ_t^{MMSE} becomes smaller. This causes the estimator to behave more conservatively by placing greater weight on the prior term $\hat{\theta}_{t|t-1}$ and reducing the influence of the uncertain innovation term $y_t - \hat{x}_t^\top \hat{\theta}_{t|t-1}$.

3.2.2 Robustness analysis and non-linear state estimation

In line with the robustness perspective of Huber and Ronchetti (2009), $\hat{\theta}_{t|t}^{\mathrm{MMSE}}$ is efficient in that it exactly recovers the Kalman Filter when covariates are deterministic (i.e., $\Omega_t = 0$). Moreover, $\hat{\theta}_{t|t}^{\mathrm{MMSE}}$ is conditionally unbiased with respect to the past observations $y_{1:t-1}$. Furthermore, among all linear estimators in the class $\mathcal{C}_t^{\mathrm{lin}}$, $\hat{\theta}_{t|t}^{\mathrm{MMSE}}$ attains the minimum variance.

As $\omega_t \to \infty$, we consider the following local asymptotic state-space model:

$$\begin{aligned} &\text{State:} && \theta_t = A\theta_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, Q_t), \\ &\text{Observation:} && \bar{y}_t = \bar{\xi}_t \theta_t, \quad \bar{\xi}_t \sim \mathcal{N}(0,1). \end{aligned} \tag{9}$$

We consider the following filtration $\bar{\mathcal{F}}_t = \sigma\{\bar{\xi}_{1:t}, \bar{y}_{1:t}\}$. Under model (9), we focus on deriving the form of state estimator and its variance, as

$$\bar{\theta}_{t|t} = \mathbb{E}\left[\theta_t \mid \bar{\mathcal{F}}_{t-1}, \bar{y}_t\right], \tag{10a}$$

$$\bar{P}_{t|t} = \mathbb{E}\left[\theta_t \mid \bar{\mathcal{F}}_{t-1}, \bar{y}_t\right]. \tag{10b}$$

Proposition 3.1 *Consider the state-space model* (9)*,*

$$\mathbb{E}\left[\theta_t \mid \bar{\mathcal{F}}_{t-1}, \bar{y}_t\right] = sign(\bar{y}_t) \frac{\int_{\mathbb{R}} g(u)\, du}{\int_{\mathbb{R}} |u|^{-1} g(u)\, du},$$

where

$$g(u) = \exp\left\{ -\frac{1}{2}\left(\frac{1}{\bar{P}^{\star}_{t|t-1}} \left(u - \bar{\theta}^{\star}_{t|t-1}\right)^2 + \frac{\bar{y}_t^2}{u^2} \right) \right\} .$$

Proposition 3.2 *(Local)*

Consider the state-space model (9)*. The density* g *defined in Proposition 3.1 is two times continuous differentiable and strictly concave (that is,* $\partial_u^2 \, g(u) \preceq 0$ *for all* $u \in \mathbb{R}$*). Define the neighbourhood*

$$\mathcal{N}_\rho = \{ \, u \in \mathbb{R} : \mid u - \bar{\theta}^{\star}_{t|t-1} \mid \leq \ \rho \, \},$$

for some $\rho > 0$*. The unique maximiser* $\bar{\theta}^{o}_{t|t}$ *of* $g(u)$ *in* $\mathcal{N}_\rho$ *equals*

$$\bar{\theta}^{o}_{t|t} = \bar{\theta}^{\star}_{t|t-1} + \frac{2\,\bar{\theta}^{\star}_{t|t-1}\bar{P}^{\star}_{t|t-1}\bar{y}_t^2}{\bar{\theta}^{\star 4}_{t|t-1} + 6\,\bar{P}^{\star}_{t|t-1}\bar{y}_t^2} + O(\rho^2).$$

The result of Proposition 3.2 leads to a non-linear form for large covariates variance. For that, assuming linearity in case of large Ω_t in the form of $\hat{\theta}_{t|t}$ is very restrictive.

Because of this, we proposed non-linear estimators that are capable even in large variance scenarios of estimating the state estimator based on new observations online.

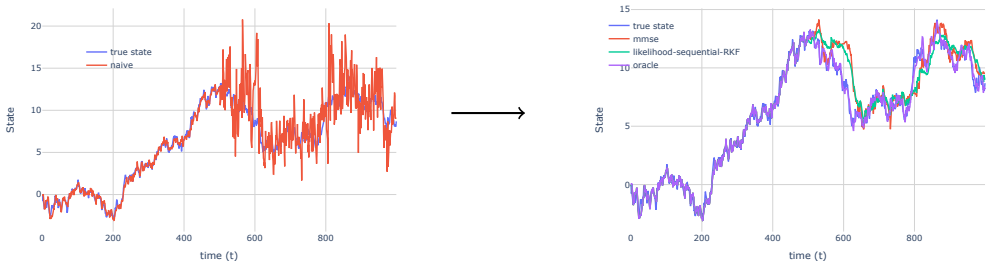


Figure 1: Left : naive estimation. Right : MMSE and Non-linear estimation.

References

Joseph de Vilmarest and Olivier Wintenberger. Viking: Variational bayesian variance tracking. *arXiv preprint arXiv:2104.10777*, 2021.

James Durbin and Siem Jan Koopman. *Time series analysis by state space methods*, volume 38. OUP Oxford, 2012.

Ludwig Fahrmeir. Posterior mode estimation by extended kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association*, 87 (418):501–509, 1992.

PJ Huber and EM Ronchetti. Robust statistics, john wiley & sons, inc. *Publication*, 2009.

Rudolph E Kalman and Richard S Bucy. New results in linear filtering and prediction theory. -, 1961.

Noboru Murata and Shun-ichi Amari. Statistical analysis of learning dynamics. *Signal Processing*, 74(1):3–28, 1999.

T Nishimura. Modeling errors in kalman filters. Technical report, -, 1970.

Yann Ollivier. Online natural gradient as a kalman filter. -, 2018.

Won-Sang Ra and Ick-Ho Whang. Stochastic robust kalman filtering for linear time-varying systems with a multiplicative measurement noise. *IFAC Proceedings Volumes*, 41(2):12546–12551, 2008.

Laura Ryan Ray and Robert F Stengel. Application of stochastic robustness to aircraft control systems. *Journal of Guidance, Control, and Dynamics*, 14(6):1251–1259, 1991.

Robert F Stengel and Laura R Ray. Technical notes and correspondence: Stochastic robustness of linear time-invariant control systems. *NASA. Langley Research Center, Joint University Program for Air Transportation Research, 1990-1991*, 1991.

Yahali Theodor and Uri Shaked. Robust discrete-time minimum-variance filtering. *IEEE Transactions on Signal Processing*, 44(2):181–189, 1996.

M. Toda and R. Patel. Bounds on estimation errors of discrete-time filters under modeling uncertainty. *IEEE Transactions on Automatic Control*, 25(6):1115–1121, 1980. doi: 10.1109/TAC.1980.1102502.

Joseph de Vilmarest. *Modèles espace-état pour la prévision de séries temporelles. Application aux marchés électriques*. PhD thesis, Sorbonne université, 2022.

Eric A Wan and Rudolph Van Der Merwe. The unscented kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*, pages 153–158. Ieee, 2000.

Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. The MIT press, 1949.

Rama K Yedavalli. Robust control of uncertain dynamic systems. *AMC*, 10:12, 2014.

Caibin Zeng, YangQuan Chen, and Qigui Yang. Robust controllability of interval fractional order linear time invariant stochastic systems. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 4047–4050, 2012. doi: 10.1109/CDC.2012.6425949.

Statistical Analysis of Economic Factors Influencing Consumer Choice by Vehicle Engine Type in the French Automotive Market

Claire Verdier[1,2], Yacouba Boubacar Maïnassara[3], Jihad Elnaboulsi[4], Camelia Goga[1] & Jean-Luc Marini[2]

[1] *Université Marie et Louis Pasteur, CNRS, LmB (UMR 6623), F-25000 Besançon, France*
[2] *Openstudio, 1 place Bellecour, F-69002 Lyon, France*
[3] *Université Polytechnique Hauts-de-France, INSA Hauts-de-France, CERAMATHS, F-59313, Valenciennes, France*
[4] *Université Marie et Louis Pasteur, CRESE (UR 3190), F-25000 Besançon, France*

cverdier@openstudio.fr, Yacouba.BoubacarMainassara@uphf.fr, jihad.elnaboulsi@univ-fcomte.fr, camelia.goga@univ-fcomte.fr, jlmarini@openstudio.fr

Abstract. The automotive industry is undergoing a major transformation driven by stringent regulations, particularly within the European Union (EU). The EU's directive to ban the sale of combustion engine vehicles by 2035 is accelerating the shift towards electric and hybrid vehicles, reshaping the industry's landscape. This regulation has a significant impact on the automotive industry, particularly in France, where more than 90% of new vehicles sold in 2018 were internal combustion engines. By examining the interplay between economic indicators and consumer preferences, the study aims to uncover the underlying dynamics driving the adoption of combustion, electric and hybrid engines. Using advanced statistical methods, including ARMA-GARCH time series models, the study examines the impact of both endogenous and exogenous variables on vehicle sales from 2010 to 2024. The results show that both internal and external variables affect passenger car registrations, with economic indicators having a different impact on each engine type.

Keywords. Adverse shocks, ARMA-GARCH Model, Automotive Industry, Consumer Behavior, Energy Economic

1 Introduction

The global industrial landscape is in a state of flux, driven by a multitude of external and internal forces. These forces are often referred to as shocks. It encompass a wide range of phenomena, from economic downturns and technological innovations to regulatory shifts and environmental pressures. For example, the COVID-19 pandemic (Li, Chen et Zhang (2017)) is a shock. Each industrial sector responds to these challenges in unique ways, adapting their strategies and outcomes to navigate the evolving landscape. Among these sectors, the automotive industry is experiencing a profound transformation, particularly due to environmental regulations and the transition to electric vehicles. This shift is not merely a technological

change but a comprehensive disruption that affects every aspect of the industry, including supply chains, consumer behavior, and market dynamics. Understanding these transformations is crucial for stakeholders to develop resilient strategies and thrive in the new era of mobility.

2 Methods

2.1 SARMAX-GARCH model

This study aims to predict passenger car sales in France by examining economic indicators that influence consumer choices. To achieve this, we propose using a SARMAX-GARCH time series model, which is an extension of the ARMA-GARCH model, well-suited for highly variable time series data, such as financial data (Francq et Zakoïan (2004), Francq et Thieu (2019)). Given the seasonality in our data, we employ a specific ARMA-GARCH model that incorporates a seasonality variable, referred to as SARMA-GARCH. Additionally, the model includes exogenous variables in the ARMA component, commonly known as ARMAX models.

Let y be an endogenous variable with y_t representing the measurement for the t-th time period, and let x_{t-1} be the vector of exogenous variable measurements recorded at the $(t-1)$-th time. The SARMAX model to predict y is defined as follows:

$$\begin{aligned} y_t =& c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_i + \epsilon_t - \sum_{i=1}^{p} \sum_{j=1}^{P} \phi_i \Phi_j y_{t-jS-i} + \sum_{j=1}^{P} \Phi_j y_{t-jS} \\ &+ \sum_{j=1}^{Q} \Theta_j \epsilon_{t-jS} - \sum_{i=1}^{q} \sum_{j=1}^{Q} \theta_i \Theta_j \epsilon_{t-jS} + \boldsymbol{\gamma}^\top \mathbf{x}_{t-1} \end{aligned} \tag{1}$$

where, p is the maximum order for non-seasonal auto-regressive coefficient, P is the maximum order for seasonal auto-regressive coefficient, q is the maximum order for non-seasonal for moving-average coefficient, Q is the maximum order for seasonal for moving-average, S is the seasonal coefficient, ϕ_i is the non-seasonal auto-regressive coefficient of order p, θ_i is the non-seasonal moving average coefficient of order q, $\boldsymbol{\gamma}$ is the vector of the unknown coefficients of the exogenous variables, Φ_i is the seasonal auto-regressive coefficient of order P with seasonality s, Θ_i is the seasonal moving average coefficient of order Q with seasonality s. Finally, ϵ_t represents the error terms which can be defined by a GARCH model:

$$\begin{aligned} \epsilon_t &= \sigma_t \eta_t, \quad \eta_t \overset{iid}{\sim} \mathcal{N}(0,1) \\ \sigma_t^2 &= \omega + \sum_{i=1}^{r} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2, \end{aligned} \tag{2}$$

where ω is a constant term, α_i are the coefficients for the lagged squared errors, β_j are the coefficients for the lagged conditional variances.

2.2 Data description

Vehicles sale data is extracted from the French government website for sustainable development statistics[1]. The study aims to The aim to forecast new vehicle sales in France by considering the economic indicators that may influence consumer choice. These indicators can include government incentives, such as the ecological bonus, or the average price of a vehicle. These data are not open-source. A brainstorming session was conducted to identify open-source data that could serve as proxies for the unavailable economic indicators and help explain consumer choices. This study adopts a data-driven methodology, leveraging open data sources to gain insights into the economic factors influencing consumer behavior. Therefore, the exogenous variables come from he French national statistical office (INSEE[2]), the Banque de France, Euorpean Central Bank and also the French government website for sustainable development statistics. The vector of exogenous variables includes other vehicles registrations like motorbike, moped and electric buggy registrations. It can serve as alternative modes of transportation to passenger cars. Furthermore, the vector of exogenous variables includes macroeconomic variables like turnover index for short-term vehicle rental, turnover index for long-term vehicle rental, new car price index, second-hand car price index, the unemployed rate by age group, the interest rate, the household debt, the Gross Disposable Income (GDI) purchasing power, the savings rate and the GDP.

3 Results

ARMA models require the stationarity of the time series. The augmented Dickey-Fyller test is therefore used to determine how many unit roots the time series contains. The stationary test is significant after removing the trend and the seasonality from *electric vehicle sales* series (see Table 1). On the other hand, for *internal combustion engine vehicle sales* series, a differentiation is necessary to make the series stationary (see Table 1).

This study examines how economic indicators affect new passenger vehicle purchases in France from 2010 to 2024, utilizing a time series model. The results show that economic indicators play a significant role in vehicle purchases, with varying impacts depending on the engine type. These insights underscore the complex interplay between economic indicators and consumer preferences within the French automotive market.

The SARIMAX-GARCH model's results highlight the significant economic factors influencing the regis- tration of electric passenger vehicles and the registration of internal combustion engine passnger vehicles in France. The interest rate, the GDP, and registrations of

[1]https://www.statistiques.developpement-durable.gouv.fr/

[2]Institut national de la statistique et des études économiques, https://www.insee.fr/fr/accueil

Without the trend	**ADF statistic**	**p-value**
electric vehicles sales	-3,283523	0,015631
internal combustion engine vehicles sales	-1,213463	0,667777

Without the trend and the seasonality	**ADF statistic**	**p-value**
electric vehicles sales	-2.53711	0.10670
internal combustion engine vehicles sales	-2.28644	0.17638

Without the trend and the seasonality First differences variables	**ADF statistic**	**p-value**
electric vehicles sales	-4.94581	0.00003
internal combustion engine vehicles sales	-5.67272	0.00000

Table 1: ADF test results

other vehicle types emerge as critical determinants of consumer to choose electric passenger vehicles. Turnover index for long-term vehicle rental and GDP are important economic indicators for the decision of consumers to buy a new internal combustion engine vehicle. These findings underscore the importance of economic conditions and market dynamics in shaping the adoption of electric vehicles.

Choosing an ARMA-GARCH model for time series analysis with shocks is a robust and flexible approach that captures both the mean and volatility dynamics of the data. This makes it particularly suitable for time series where shocks are common and have a significant impact on the series.

Bibliographie

Francq, C. and Thieu, L. Q. (2019). QML inference for volotility models with covariates. Econometric Theory, 35(1):37–72.

Francq, C., Zakoian, J. M. (2004). Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. Bernoulli, 10(4), 605-637.

Li, S., Chen, H., and Zhang, G. (2017). Comparison of the Short-Term Forecasting Accuracy on Battery Electric Vehicle between Modified Bass and Lotka-Volterra Model: A Case Study of China. Journal of Advanced Transportation, 2017(1), 7801837.

Stechemesser, A., Koch, N., Mark, E., Dilger, E., Klösel, P., Menicacci, L., ... and Wenzel, A. (2024). Climate policies that achieved major emission reductions: Global evidence from two decades. Science, 385(6711), 884-892.

Séries temporelles – Théorie

Sur les bruits blancs faibles modulés par une chaîne de Markov cachée

Armel Bra[*1], **Yacouba Boubacar Maïnassara**[†2], and **Landy Rabehasaina**[‡3]

[2]Université Polytechnique Hauts-de-France, INSA Hauts-de-France, CERAMATHS-Laboratoire de Matériaux, Céramiques et de Mathématiques, F-59313 Valenciennes, France.

[1,3]Université Marie et Louis Pasteur, Laboratoire de mathématiques de Besançon, 16 route de Gray, 25030 Besançon cedex, France.

Résumé. Dans ce document, nous présentons les propriétés asymptotiques de l'estimateur du maximum de vraisemblance d'un processus appartenant à la famille des chaînes de Markov cachées (HMM) intégrant des changements de régime markoviens où les erreurs sont non corrélées mais pas nécessairement indépendantes, avec l'hypothèse que les régimes ne sont pas directement observables. L'assouplissement des hypothèses concernant la non-indépendance des erreurs et la non-observabilité directe des régimes élargit significativement l'applicabilité de cette classe de modèles à changements de régimes. Nous donnons des conditions nécessaires pour prouver la consistance et la normalité asymptotique de l'estimateur du maximum de vraisemblance.

Mots-clés. Processus non linéaire, processus stationnaire, chaîne de Markov cachée, changement de régime, vraisemblance maximum, cohérence, normalité asymptotique.

Abstract. In this document, we present the asymptotic properties of the maximum likelihood estimator for a process belonging to the family of hidden Markov chains (HMM) incorporating Markovian regime changes where the errors are uncorrelated but not necessarily independent, under the assumption that the regimes are not directly observable. Relaxing the assumptions concerning the non-independence of the errors and the direct non-observability of the regimes significantly broadens the applicability of this class of regime-switching models. We provide necessary conditions to prove the consistency and asymptotic normality of the maximum likelihood estimator.

Keywords. Non-linear process, stationary process, hidden Markov chain, regime change, maximum likelihood, consistency, asymptotic normality.

1 Introduction

Les séries chronologiques présentent souvent des changements brusques dans leur dynamique, rendant les modèles linéaires classiques tels que les modèles autorégressifs à moyenne mobile (ARMA pour faire court) ou l'hétéroscédasticité conditionnelle autorégressive généralisée (GARCH pour faire court) inadéquats pour décrire certains phénomènes observés. Par exemple, les indicateurs financiers ou économiques peuvent passer soudainement de phases à faible volatilité calmes à des phases à haute volatilité turbulentes. Dans Hamilton (1989), ces changements de volatilité ont été traités comme des changements d'état (ou de régime) régis par une chaîne de Markov. Dans de tels cas, les hypothèses standard de stationnarité et d'homoscédasticité ne tiennent plus, nécessitant

[*]kja.bra@univ-fcomte.fr
[†]Yacouba.BoubacarMainassara@uphf.fr
[‡]landy.rabehasaina@univ-fcomte.fr

de nouvelles stratégies de modélisation qui capturent les comportements dépendants des régimes. Une solution naturelle consiste à permettre aux paramètres régissant la série chronologique (par exemple, la variance ou les coefficients autorégressifs) de varier en fonction d'un état latent (non observé) qui évolue dans le temps. Plus précisément, on peut introduire une *chaîne de Markov cachée*, puisque le moment des changements de régime n'est pas directement observé, et cette chaîne détermine le régime ou l'état à chaque instant t.

Dans cet article, nous nous concentrons sur un modèle à structure multiplicative qui dépend à la fois d'une chaîne de Markov cachée et d'un bruit exogène, supposé non corrélé mais pas nécessairement indépendant. Ce type de bruit est souvent appelé bruit blanc faible, contrairement au bruit blanc fort, qui est typiquement indépendant et identiquement distribué (voir Romano & Thombs (1996) pour un exemple de bruit blanc faible). Ces modèles peuvent être interprétés comme des mélanges markoviens de systèmes dynamiques, les plaçant dans la classe des modèles de Markov cachés (HMM pour faire court); voir, par exemple, Francq & Roussignol (1997) et Hamilton (1988). On pourrait envisager d'ajuster de tels modèles en utilisant des approches d'hétéroscédasticité conditionnelle autorégressive (ARCH pour faire court) introduites par Engle (1982) (voir également Gourieroux (1992) pour plus de détails). Cependant, les modèles ARCH sont principalement conçus pour capturer la volatilité conditionnelle induite par les valeurs passées; ils ne prennent pas facilement en compte les changements de régime soudains dans la dynamique des séries chronologiques. Dans notre cadre, ces changements soudains sont modélisés naturellement comme des changements d'état dans une chaîne de Markov cachée.

Les recherches récentes ont encore fait progresser ce domaine. Le calcul direct de la vraisemblance dans les modèles de Markov cachés est complexe, mais plusieurs méthodes y répondent, y compris l'algorithme avant-arrière proposé par Baum *et al.* (1972) et l'algorithme EM (Expectation-Maximization) pour l'estimation du maximum de vraisemblance décrit par Baum *et al.* (1970) et Dempster *et al.* (1977). Les propriétés statistiques de ces estimateurs ont également été étudiées en profondeur. Par exemple, Leroux (1992) ont établi des conditions pour la consistance de l'estimateur du maximum de vraisemblance dans les HMM généraux, tandis que Rydén (1994) ont introduit une approche de pseudo-maximum de vraisemblance pour prouver à la fois la consistance et la normalité asymptotique. De même, Francq & Roussignol (1997) ont démontré la consistance et la normalité asymptotique sous des hypothèses de mélange spécifiques sur le bruit (voir également Douc *et al.* (2004)).

La plupart des résultats existants reposent sur une hypothèse d'indépendance et de distribution identique (i.i.d.) pour le bruit. En revanche, lorsque le bruit est seulement un bruit blanc faible et que le modèle inclut une chaîne de Markov cachée, la littérature s'est jusqu'à présent principalement concentrée sur les propriétés de stationnarité (notamment, Francq & Zakoıan (2001) les ont étudiées dans un contexte multivarié). L'objectif principal de ce travail est d'étendre les résultats de Francq & Roussignol (1997), initialement prouvés dans un cadre de bruit i.i.d., au cas d'un bruit blanc faible. Nous employons un algorithme pseudo-EM pour l'estimation des paramètres dans notre cadre et établissons par la suite à la fois la consistance et la normalité asymptotique du pseudo-estimateur.

2 Modèle et hypothèses

Nous considérons un processus stationnaire $(\epsilon_t)_{t\in\mathbb{Z}}$ donné par

$$\epsilon_t = f^0(\Delta_t)\eta_t \qquad \forall t \in \mathbb{Z}, \tag{1}$$

où $(\eta_t)_{t\in\mathbb{Z}}$ est un bruit blanc faible vérifiant $\mathbb{E}(\eta_t\eta_{t'}) = \sigma^2\mathbb{1}_{\{t=t'\}}$, $(\Delta_t)_{t\in\mathbb{Z}}$ une chaîne de Markov cachée prenant des valeurs dans $\mathcal{E} := \{1, \cdots, d\}$ et f^0 une fonction de $\mathcal{E}$ à valeurs dans $\mathbb{R}_+^*$. De plus, nous supposons que

A.1 La chaîne de Markov $(\Delta_t)_{t\in\mathbb{Z}}$ est strictement stationnaire, irréductible et apériodique,

A.2 Les processus $(\Delta_t)_{t\in\mathbb{Z}}$ et $(\eta_t)_{t\in\mathbb{Z}}$ sont indépendants,

A.3 Le processus $(\eta_t)_{t\in\mathbb{Z}}$ est un processus ergodique,

A.4 Le processus $(\eta_t)_{t\in\mathbb{Z}}$ est distribué avec une moyenne nulle et une variance σ^2,

A.5 Étant donnée la chaîne de Markov $(\Delta_t)_{t\in\mathbb{Z}}$, le processus $(\epsilon_t)_{t\in\mathbb{Z}}$ possède une densité par rapport à la mesure de Lebesgue.

Soit $\pi_i := \mathbb{P}(\Delta_t = i), 1 \leq i \leq d$ la distribution stationnaire et $p^0_{i,j} := \mathbb{P}(\Delta_t = j \mid \Delta_{t-1} = i), 1 \leq i, j \leq d$ les probabilités de transition. Sans perte de généralité, nous supposons également que $\sigma = 1$.

Afin de mesurer la dépendance temporelle des processus $(\eta_t)_{t\in\mathbb{Z}}$ et $(\Delta_t)_{t\in\mathbb{Z}}$, nous définissons les coefficients de mélange forts $(\alpha_{\mathcal{Z}}(h))_{h\in\mathbb{N}^\star}$, qui sont indépendants de $t \in \mathbb{Z}$, pour un processus stationnaire $(\mathcal{Z}_t)_{t\in\mathbb{Z}}$ comme suit

$$\alpha_{\mathcal{Z}}(h) := \sup_{A\in\mathcal{F}^t_{-\infty}, B\in\mathcal{F}^\infty_{t+h}} \big|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)\big|, \tag{2}$$

où $\mathcal{F}^t_{-\infty}$ et $\mathcal{F}^\infty_{t+h}$ désignent les σ-algèbres générées par $\{Z_u : u \leq t\}$ et $\{Z_u : u \geq t+h\}$, respectivement.
Nous faisons donc l'hypothèse suivante

A.6 Les processus $(\eta_t)_{t\in\mathbb{Z}}$ et $(\Delta_t)_{t\in\mathbb{Z}}$ satisfont $\displaystyle\sum_{h=0}^{\infty} \alpha_\eta(h)^{\frac{1}{2}} < \infty$ et $\displaystyle\sum_{h=0}^{\infty} \alpha_\Delta(h)^{\frac{1}{2}} < \infty$,

où $\alpha_\eta(\cdot)$ (respectivement $\alpha_\Delta(\cdot)$) désignent les coefficients de mélange des processus $(\eta_t)_{t\in\mathbb{Z}}$ (respectivement $(\Delta_t)_{t\in\mathbb{Z}}$) tels que définis dans l'Équation (2).

En notant que sous l'hypothèse précédente **A.6** et l'indépendance entre $(\Delta_t)_{t\in\mathbb{Z}}$ et $(\eta_t)_{t\in\mathbb{Z}}$, il découle de (Bradley, 2005, Theorem 5.1, p. 126) que le processus $(\epsilon_t)_{t\in\mathbb{Z}}$ est fortement mélangé et que les coefficients de mélange $(\alpha_\epsilon(h))_{h\in\mathbb{Z}}$ de $(\epsilon_t)_{t\in\mathbb{Z}}$ satisfont $\alpha_\epsilon(h) \leq \alpha_\Delta(h) + \alpha_\eta(h)$.

A.7 Pour tout $\theta \in \Theta$, if $h_\theta(\epsilon_{t-m+1}, \cdots, \epsilon_t) = h_{\theta_0}(\epsilon_{t-m+1}, \cdots, \epsilon_t)\mathbb{P}_{\theta_0} - a.s$ alors $\theta = \theta_0$.

3 Estimation des paramètres

Dans notre contexte, le paramètre inconnu est $\theta_0 := ((p^0_{i,j})_{1\leq i\leq d,\ 2\leq j\leq d}, (f^0(s))_{1\leq s\leq d})$ appartenant à l'ensemble $\Theta := \Big\{((p_{i,j})_{1\leq i\leq d,\ 2\leq j\leq d}, (f(s))_{1\leq s\leq d}) \in]0,1[^{d(d-1)}\times]0,\infty[^d, \displaystyle\sum_{j=2}^{d} p_{i,j}\mathbb{1}_{1\leq i\leq d} \leq 1\Big\}$.

Soit m un entier supérieur ou égal à 2. Supposons que l'on dispose des observations $\{x_1, \ldots, x_n\}$, tandis que la chaîne de Markov $(\Delta_t)_{t\in\mathbb{Z}}$ reste non observée. Nous considérons la pseudo-vraisemblance suivante

$$\forall\ \theta \in \Theta,\ L_{m,n}(x_1, \cdots, x_n; \theta) := \prod_{t=m}^{n} h_\theta(x_{t-m+1}, \cdots, x_t) \tag{3}$$

où

$$h_\theta(x_1, \cdots, x_m) = \sum_{(e_1\cdots,e_m)} \pi_{e_1} \left\{\prod_{j=2}^{m} p_{e_{j-1},e_j}\right\} \left\{\prod_{t=1}^{m} \zeta(x_t|\Delta_t = e_t)\right\}, \tag{4}$$

avec $\zeta(x_t|\Delta_t = e_t)$ la densité de ϵ_t sachant $\Delta_t = e_t$.

Un estimateur de θ_0 est défini par $\hat{\theta}_{m,n}$, où

$$\hat{\theta}_{m,n} \ = \ \arg\max_{\theta\in\Theta} L_{m,n}\big(x_1, \ldots, x_n; \theta\big).$$

4 Propriétés asymptotique

Dans cette section, nous étudions les propriétés asymptotiques de l'estimateur $\hat{\theta}_{m,n}$. On pose

$$I_n(\theta) := \mathrm{Var}\left(\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\log L_{m,n}(\theta)\right) = \frac{1}{n}\mathbb{E}\left\{\frac{\partial}{\partial\theta'}\log L_{m,n}(\theta)\frac{\partial}{\partial\theta}\log L_{m,n}(\theta)\right\}$$

et

$$J_n(\theta) := -\frac{1}{n}\mathbb{E}\frac{\partial^2}{\partial\theta'\partial\theta}\log L_{m,n}(\theta) = \frac{n-m+1}{n}\,\mathrm{Var}\left\{\frac{\partial}{\partial\theta}\log h_\theta(\epsilon_{t-w+1},\cdots,\epsilon_t)\right\}.$$

Théorème 1. *Soit $m \in \mathbb{N}$ tel que $m \geq 2$ et $\Theta^\star$ un sous-ensemble compact de Θ contenant le véritable paramètre θ_0. Alors, toute suite $(\hat{\theta}_{m,n})_{n\in\mathbb{N},\, n\geq m}$ vérifiant*

$$L_{m,n}(x_1,\cdots,x_n;\hat{\theta}_{m,n}) = \sup_{\theta\in\Theta^\star} L_{m,n}(x_1,\cdots,x_n;\theta)$$

converge presque sûrement vers θ_0 lorsque n tend vers l'infini.

Théorème 2. *Pour tout $m \in \mathbb{N}$ tel que $m \geq 2$, le vecteur aléatoire $\sqrt{n}(\hat{\theta}_{m,n}-\theta_0)$ converge en distribution vers une distribution normale de moyenne zéro et de matrice de covariance $J(\theta_0)^{-1}I(\theta_0)J(\theta_0)^{-1}$.*

Bibliography

Baum, Leonard E, Petrie, Ted, Soules, George, & Weiss, Norman. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, **41**(1), 164–171.

Baum, Leonard E, *et al.* 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, **3**(1), 1–8.

Bradley, Richard C. 2005. Basic properties of strong mixing conditions. A survey and some open questions.

Dempster, Arthur P, Laird, Nan M, & Rubin, Donald B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, **39**(1), 1–22.

Douc, Randal, Moulines, Eric, & Rydén, Tobias. 2004. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime.

Engle, Robert F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the econometric society*, 987–1007.

Francq, Christian, & Roussignol, Michel. 1997. On white noises driven by hidden Markov chains. *Journal of Time Series Analysis*, **18**(6), 553–578.

Francq, Christian, & Zakoıan, J-M. 2001. Stationarity of multivariate Markov–switching ARMA models. *Journal of Econometrics*, **102**(2), 339–364.

Gourieroux, Christian. 1992. *Modèles ARCH et applications financières*. Economica.

Hamilton, James D. 1988. Rational-expectations econometric analysis of changes in regime: An investigation of the term structure of interest rates. *Journal of Economic Dynamics and Control*, **12**(2-3), 385–423.

Hamilton, James D. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, 357–384.

Leroux, Brian G. 1992. Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, **40**(1), 127–143.

Romano, Joseph P, & Thombs, Lori A. 1996. Inference for autocorrelations under weak assumptions. *Journal of the American Statistical Association*, **91**(434), 590–600.

Rydén, Tobias. 1994. Consistent and asymptotically normal parameter estimates for hidden Markov models. *The Annals of Statistics*, 1884–1895.

splineOP: Quadratic Spline Compression for Univariate Time Series

Nicolás Cecchi [1] , Laurent Oudre [2] , Vincent Runge[3] & Charles Truong [4]

[1 2 4] *Université Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli, France*
nicolas.cecchi1@ens-paris-saclay.fr,laurent.oudre@ens-paris-saclay.fr,
charles.truong@ens-paris-saclay.fr
[3] *Université Paris-Saclay, CNRS, Univ Evry, Laboratoire de Mathématiques et Modélisation d'Evry, France, vincent.runge@univ-evry.fr*

Abstract. The detection of changes in time series is a widely studied topic. In many settings, e.g. in behavioral neuroscience, time series are smooth trajectories representing the movements of patients' body. Our objective is to approximate those time series by differentiable functions. More precisely, we use a piecewise polynomial approximation with continuity constraints and only a few changes. The differentiability constraints make the optimization problem hard to solve. We propose a dynamic programming recursion to find the optimal solution and an approximating algorithm by locally discretizing the space of admissible values at the change-points and of initial first derivatives. We validate its performance over quadratic splines issued from a real medical data-set with varying sample sizes and signal-to-noise ratios.

Keywords. time-series, change-point detection, dynamic programming, splines

1 Introduction

Change point detection has become a standard task in time series analysis. Numerous algorithms exist to accurately and efficiently estimate the number and locations of changes in time series, many of them based on dynamic programming, proposed by Bellman (1954) or binary segmentation, by Vostrikova (1981). Among these, the detection of changes in mean has been extensively studied, as in Page (1954), Srivastava and Sen (1975) and Fearnhead and Rigaill (2020). However, detecting changes when dependencies across consecutive segments exist remains challenging. A notable example is change in slope detection under continuity constraints, studied by Maidstone, Fearnhead, and Letchford (2017), Runge, Pascucci and de Boishebert (2020) and Fearnhead and Grose (2024).

In this work, we consider the next challenging task of recovering a piecewise quadratic function. This problem is of practical use in the analysis of smooth trajectory data. We propose an exact and an approximate solution with a dynamic programming algorithm demonstrating promising results. Notably, it makes for an excellent change-point selector, making it useful for high-quality compression of large datasets.

2 Detecting changes in K_2-splines

2.1 Problem statement

A real-valued function p on $[0, 1]$, composed of K polynomials of up to degree 2, is called a K_2-spline if it is continuous and has continuous derivatives in its domain. Change points are those $\tau \in (0, 1)$ where the second derivative is discontinuous, i.e. $p''(\tau^-) \neq p''(\tau^+)$.

Consider data $y = \{y_1, y_2, ...y_T\}$ drawn at successive positions $t_i \in [0, 1)$ from a K_2-spline f with Gaussian noise. We search for the best element $\hat{f}$ in S_{K_2} minimizing the following quantity:

$$\hat{f} = \underset{f \in S_{K_2}}{argmin} \Big\{ \sum_{i=1}^{T} (y_i - f(t_i))^2 + \beta I(f''(t_i^-) \neq f''(t_i^+)) \Big\}, \tag{1}$$

where $\beta > 0$ is a given penalty value, I the indicator function and f is optimized over S_{K_2}, the set of K_2-splines. In the above objective, the L_0 penalty promotes sparsity of changes.

2.2 Solution

We propose an exact dynamic programming solution, which is computationally intractable. We hence propose an alternative approximating solution, by discretizing the space of parameters and the following recursion formula that can be solved with a dynamic-programming algorithm of complexity $O(N^2T^2)$:

$$Q_t(p, v_{p,t}) = \min_{\substack{\tilde{p} \in P_s \\ 0 \leq s < t}} \Big\{ Q_s\Big(\tilde{p}, v_{\tilde{p},s}\Big) + C_{s:t}(\tilde{p}, p, v_{p,t}) + \beta \Big\}, \; v_{p,t} = 2\frac{p - \tilde{p}}{t - s} - v_{\tilde{p},s}, \; t \in \{1, \ldots, T\}, \tag{2}$$

with initial condition $Q_0(p, v_0) = -\beta$ for all $p \in P_t$ and v_0 a well-chosen initial speed.

2.3 Simulation study

We perform numerical tests on signals of different lengths and with different signal-to-noise ratios, issued from a real medical dataset. We peform simulations with varying access to 'oracle' knowledge and heuristics. The results are promising, as the polynomials being fit out of the predicted change-points can approximate accurately the underlying data generating function.

3 Conclusions

In this work we studied the problem of finding changes in the second derivative of K_2-splines and developed and efficient algorithm to estimate their locations. Experiments on numerical data shows promising results. Which is motivating for further work such as studying higher order polynomials, other heuristics or finding theoretical bounds for the error being made.

Bibliography

Bellman, R. (1954), The Theory of Dynamic Programming, *Bulletin of the American Mathematical Society*, 60, pp. 503-515

Vostrikova, L. Y. (1981), Detecting "disorder" in Multidimensional Random Processes, *Dokl. Akad. Nauk SSSR*, Volume 259, Number 2, pp. 270–274

Page, E.S. (1954), Continuous Inspection Schemes, *Biometrika*, Vol. 41, No. 1/2, pp. 100-115

Sen A. and Srivastava, M. S. (1975), On Tests for Detecting Change in Mean, *The Annals of Statistics* Vol. 3, No. 1, pp. 98-108

Fearnhead, P. and Rigaill, G. (2020), Relating and Comparing Methods for Detecting Changes in Mean, *Stat*

Maidstone, R., Fearnhead, P., and Letchford, A. (2017). Detecting Changes in Slope with an L_0 Penalty. arXiv preprint arXiv:1701.01672.

Runge, V., Pascucci, M., and de Boishebert, N. D. (2020). Change-in-slope Optimal Partitioning Algorithm in a Finite-size Parameter Space. arXiv preprint arXiv:2012.11573.

Fearnhead, P., and Grose, D. (2024). cpop: Detecting Changes in Piecewise-linear Signals. Journal of Statistical Software, 109, 1-30.

Échangeabilité et Facteurs Faibles

Marc Hallin

Université libre de Bruxelles, Belgique, mhallin@ulb.be

Résumé. La littérature économétrique montre beaucoup d'intérêt pour la notion de facteur faible dans les modéles à facteurs, qu'ils soient de nature statique ou dynamique. Des tests ont même été proposés pour l'hypothése nulle d'un facteur fort (divergence linéaire de la valeur propre associée à ce facteur) par rapport à une contre-hypothèse de facteur faible (divergence sublinéaire de cette valeur propre). Nous montrons que la notion de facteur faible, dans un panel de dimension nxT est incompatible avec l'échangeabilité du processus observé, hypothése sous laquelle l'ordre dans lequel sont rangées les n séries observées, est sans impact sur les résultats; cet ordre, en effet, est généralement arbitraire.

Basé sur une collaboration avec Matteo Barigozzi (Bologne)

Mots-clés. Modèles à facteurs; facteurs faibles; échangeabilité.

Abstract. Many authors have discussed the presence of rate-weak factors in factor models. Tests even have been proposed for the null hypothesis of a strong factor against the alternative of a weak one (linear versus sublinear divergence of the corresponding eigenvalues). We show that this existence of rate-weak factors is incompatible with the natural assumption that the cross-sectional ordering of the panel, which usually is arbitrary, should remain irrelevant.

Based on joint work with Matteo Barigozzi (Bologna).

Keywords. Characteristic function; total independence; independent component model; rank test.

*Marc Hallin gratefully acknowledges the support of the Czech Science Foundation grant GA24-10078S.

Les modèles à facteurs ont pour objet l'analyse et la prédiction des grands panels de séries temporelles. Ceux-ci se présentent sous la forme de réalisations finies (n séries, indexées $i = 1, \ldots, n$, observées aux dates $t = 1, \ldots, T$; n "grand") de processus stochastiques (à variance finie; par simplicité, toutes les variables sont supposées de moyenne nulle) de la forme

$$\{X_{it} | \, i \in \mathbb{N}, \, t \in \mathbb{Z}\}.$$

Cette analyse se fonde sur une décomposition

$$X_{it} = \chi_{it} + \xi_{it} \qquad i \in \mathbb{N} \; t \in \mathbb{Z}$$

en une composante *commune* χ_{it} et une composante *idiosyncratique* ξ_{it}. La composante commune rend compte des corrélations croisées *pervasives* liées à des chocs ou des facteurs communs, tandis que les corrélations croisées idiosyncratiques sont *non-pervasives*. Cette "pervasivité" est liée au comportement, explosif ou borné pour $n \to \infty$, de valeurs propres: les valeurs propres de la matrice de covariance des n premières séries dans le cas des modèles dits "statiques", les valeurs propres de leur matrice de densité spectrale dans le cas des modèles dits "dynamiques". A chaque valeur propre explosive est associé un facteur (cas statisue) ou un choc commun (cas dynamique).

La littérature économétrique, dans ce contexte, montre beaucoup d'intérêt pour la notion de facteur faible dans les modéles à facteurs, qu'ils soient de nature statique ou dynamique. Des tests ont même été proposés pour l'hypothése nulle d'un facteur fort (divergence linéaire de la valeur propre associée à ce facteur) par rapport à une contre-hypothèse de facteur faible (divergence sublinéaire de cette valeur propre).

Une propriété trop souvent négligée des panels étudiés au moyen de modèles à facteurs, toutefois, est le caractère entièrement arbitraire de l'ordre dans lequel sont rangées les n séries observées—un ordre qui, par conséquent, ne devrait pas avoir d'impact sur l'analyse et la prédiction. Cette indépendance par rapport à l'attribution aux séries observées des labels $i = 1, \ldots, n$ se traduit naturellement en une hypothèse d'*échangeabilité transversale* (au second ordre) du processus $\{X_{it}\}$, qui vient compléter l'habituelle propriété de stationarité (au second ordre):

Définition 1. Le processus $\mathbf{X} := \{X_{it}\}$ appartient à la classe $\mathcal{C}$ des processus stationnaires au second ordre: pour tout i, $i' \in \mathbb{N}$, tout t, $t' \in \mathbb{Z}$, et tout $h \in \mathbb{Z}$,

$$\mathrm{E}\left[X_{it} X_{i't'}\right] = \mathrm{E}\left[X_{i,t+h} X_{i',t'+h}\right].$$

Définition 2. Le processus $\mathbf{X} \in \mathcal{C}$ est échangeable au second ordre: pour tout $k \in \mathbb{N}$, tout k-tuple $i_1, \ldots, i_k$, et toute permutation π dans l'ensemble Π_k des $k!$ permutations de $(1, \ldots, k)$, les moments du second ordre des processus k-dimensionnels

$$\left\{(X_{i_1 t}, \ldots, X_{i_k t}) | \, t \in \mathbb{Z}\right\} \quad \text{et} \quad \left\{(X_{i_{\pi(1)} t}, \ldots, X_{i_{\pi(k)} t}) | \, t \in \mathbb{Z}\right\}$$

coïncident.

Cet exposé a pour but de montrer que la notion de facteur faible (mais aussi celle de facteur super-linéaire) est incompatible avec ces hypothèses.

Références.

Barigozzi, M. and Hallin, M. (2024). The Dynamic, the Static, and the Weak. *Journal of Time Series Analysis*, special issue in honor of Hira Koul, to appear; available at `http://arxiv.org/abs/2407.10653`.

On self-excited threshold bivariate Poisson integer-valued $GARCH$ models

Mohamed SADOUN [1] & Mehdi FRAHI [2] & Roufaida SOUAKRI [3]

[1] *RECITS Laboratory, Faculty of Mathematics, University of Sciences and Technology Houari Boumediene, Algeria, mo-hamedsadoun@outlook.fr*

[2] *Faculty of Mathematics, University of Sciences and Technology Houari Boumediene, Algeria, frahimehdi7@gmail.com*

[3] *Faculty of Mathematics, University of Sciences and Technology Houari Boumediene, Algeria, roufaidasekiou11@gmail.com*

Résumé. Cette contribution propose un modèle autorégressif à seuil bivarié à valeurs entières, à régimes multiples ($K \geq 2$), basé sur une distribution de Poisson bivariée conditionnelle (ci-après désignée par $SET - BPINGARCH\,(K; 1, 1)$). Ce modèle sera utile pour analyser le nombre bivarié de certaines arrivées dans un intervalle de temps fixe avec un comportement de non-linéarité. Tout d'abord, nous donnons la formulation du modèle en utilisant le cadre $INGARCH$, en particulier pour la distribution de Poisson bivariée avec réduction trivariée et la distribution de Poisson bivariée avec un facteur multiplicatif. Ensuite, nous fournissons deux procédures d'inférence statistique pour les cas de seuil connus et inconnus, à savoir : deux méthodes d'estimation incluant la méthode des moindres carrés ordinaires (OLS), et la méthode du maximum de vraisemblance conditionnelle (CML). Enfin, les performances des procédures d'inférence obtenues seront évaluées via une étude intensive de simulation et une application sur des données réelles.

Mots-clés. Poisson bivarié $INGARCH$, modèle à seuil auto-excité $BP - INGARCH$, réduction trivariée, estimateurs (OLS), estimateurs (CML).

Abstract. This contribution proposes a bivariate threshold integer-valued autoregressive model, with multiple regimes ($K \geq 2$), based on Conditional bivariate Poisson distribution (hereafter referred to as $SET - BPINGARCH\,(K; 1, 1)$). This model will be useful for analyzing the bivariate number of certain arrivals in a fixed time interval with nonlinearity behavior. First, we give the formulation of the model using the $INGARCH$ framework, particularly for the Bivariate Poisson distribution with trivariate reduction and the Bivariate Poisson distribution with a multiplicative factor. Second, we provide two statistical inference procedures for both known and unknown threshold cases, namely: two estimation methods including the ordinary least squares (OLS) method, and the conditional maximum likelihood (CML) method. Finally, the performance of the obtained inference procedures will be evaluated via an intensive simulation study and its application to real data.

Keywords. Bivariate Poisson $INGARCH$, Self-Excited Threshold $BP - INGARCH$ model, trivariate reduction, (OLS) and (CML) estimators.

1 Introduction and motivations

Numerous count time series encountered in various fields reveal strong nonlinearity characteristics that cannot be captured by the standard linear count time series models. This nonlinearity fact is often considered to be a piecewise linear "threshold" and was emphasized in several works in the literature, we have: Thyregod *et al* (1999), Yang *et al* (2018), Bentarzi and Sadoun (2021), and many others. These compelling aspects of the threshold model motivated us to extend it to multivariate cases through a novel formulation that incorporates multiple thresholds based on the conditional bivariate Poisson distribution. Multivariate count time series data with correlated counts observed at different time points are prevalent in various practical fields. Notable applications include epidemiology, finance, geophysics, criminology, and marketing (Karlis, 2016). Numerous approaches exist for modeling count time series, leading to a diverse range of models. Several authors have contributed to bivariate and multivariate modeling, including Liu (2012), Cui (2018), Fokianos (2020, 2021), and Lee (2023). In our work, we shall focus on a conditional multivariate distribution-based model of $INGARCH$ type. The $INGARCH$ class of models is generally defined by a conditional distribution. Given a series of observations $\mathbf{Y}_t$ generated by a specific process, each observation at time t follows a certain distribution conditional on its past. The link function of this distribution is modeled through an $ARMA$ process. A multivariate $INGARCH$ model with orders p and q can generally be given by:

$$\begin{aligned} \mathbf{Y}_t|\mathcal{F}_{t-1} &\sim \mathcal{L}(\cdot) \\ \sigma_t &= \omega + \sum_{i=1}^{p} B_i \mathbf{Y}_{t-i} + \sum_{j=1}^{q} A_j \sigma_{t-j} \end{aligned} \tag{1}$$

where $\mathbf{Y}_t$ is a d-dimensional vector of observations at time t, $\mathcal{L}$ is a discrete distribution and σ a link function, ω a d-dimensional vector, B_i and A_j $((i,j) \in \{1,\ldots,p\} \times \{1,\ldots,q\})$ square matrices with size $d \times d$. In our case, $\mathcal{L}(.) \rightsquigarrow \mathcal{P}(.)$ involving an $INGARCH$ bivariate Poisson model noted ($BP-INGARCH$ model). We consider two types of bivariate Poisson distributions: the trivariate reduction type (Liu, 2012) and the bivariate Poisson distribution with a multiplicative factor (Lakshminarayana *et al*, 1999). Indeed, the bivariate Poisson distributions with trivariate reduction $B\mathcal{P}^*(\lambda_1, \lambda_2, \phi)$ is given as follows:

$$\begin{aligned} (Y_1, Y_2)' &\rightsquigarrow B\mathcal{P}^*(\lambda_1, \lambda_2, \phi) \\ Y_1 &= X_1 + X_3 \\ Y_2 &= X_2 + X_3 \\ X_1 &\rightsquigarrow \mathcal{P}(\lambda_1 - \phi);\ X_1 \rightsquigarrow \mathcal{P}(\lambda_2 - \phi);\ X_3 \rightsquigarrow \mathcal{P}(\phi) \end{aligned}$$

With a joint mass function defined by:

$$P(Y_1 = y_1, Y_2 = y_2) = e^{-(\lambda_1+\lambda_2-\phi)} \frac{(\lambda_1-\phi)^{y_1}(\lambda_2-\phi)^{y_2}}{y_1! y_2!} \sum_{k=0}^{\min(y_1,y_2)} C_{y_1}^k C_{y_2}^k k! \left(\frac{\phi}{(\lambda_1-\phi)(\lambda_2-\phi)}\right)^k$$

where y_1, y_2 are non-negative integers and $\lambda_1 - \phi$, $\lambda_2 - \phi$ and ϕ are positive parameters, $\phi \in [0, \min(\lambda_1, \lambda_2)]$.

Whereas, the bivariate Poisson distributions with trivariate reduction $B\mathcal{P}(\lambda_1, \lambda_2, \phi)$ is given by the following joint mass function:

$$P(Y_1 = y_1, Y_2 = y_2) = \frac{\lambda_1^{y_1}\lambda_2^{y_2}}{y_1!y_2!}e^{-(\lambda_1+\lambda_2)}\left[1 + \delta\left(e^{-y_1} - e^{-c\lambda_1}\right)\left(e^{-y2} - e^{-c\lambda_2}\right)\right]$$

where $c = 1-e^{-1}$, and λ_1, λ_2 are positive parameters. In this case, the margianl distributions of Y_1 and Y_2 follow a Poisson distribution with parameters λ_1 and λ_2 respectively. Using a bivariate Poisson distribution with a multiplicative factor allows us to capture positive as well as negative correlations, inherent to several data in practice. While the $BP-INGARCH$ model represents a significant improvement over previous bivariate models (Liu, 2012; Fokianos, 2022; Armillotta, 2023; Liu, 2023), certain limitations remain. One key aspect of time series data is linearity, and thus, we focus on exploring non-linearity in modeling. Expanding the model to incorporate novel regime-switching formulations allows us to tackle this challenge using threshold models. We will present estimation methods, conduct simulation studies, and illustrate the approach with real data examples.

The present work is organized as follows: In section 2, we define the new $(SET - BPINGARCH)$ model and supply some preliminary results. In section 3, we first address where the thresholds are known, the problem of estimation in a parametrical framework by considering two different methods, namely: the ordinary least squares (OLS) method, and the conditional maximum likelihood (CML) method. Next, we deal with the case where the thresholds are unknown basing on some sweeping searching approaches ($SMLS$ and SML) methods. In section 4, we present some simulation results and an application on the data set of the number of daytime and nighttime road accidents, while showing the superiority of our $SET - BPINGARCH$ model over some competitor models.

2 Self-excited Threshold BP-INGARCH Model

Let $Z_t = (Z_{t,1}, Z_{t,2})^T$ represent the bivariate time series at t. The general $SET - BP - INGARCH(K, 1, 1)$ model with K regimes is defined as follows:

$$Z_t = \sum_{\kappa=1}^{K} Y_{\kappa,t}\mathbb{I}_{S(t)=\kappa}. \tag{2}$$

where:

- $\mathbb{I}_{S(t)=\kappa}$ is the indicator function, equal to 1 if $S(t) = \kappa$ and 0 otherwise.
- S_t is a sequence of random variables (not independent of $Z_{t-i}, i \geq 1$) indicating the current regime.
- Z_t encompasses all observations, partitioned into $Y_{\kappa,t}$ based on the regime.

Each regime $\kappa \in 1, 2, ..., K$ follows either a $BP^{*}-INGARCH(1,1)$ or a $BP-INGARCH(1,1)$ process:

$$\begin{aligned} Y_{\kappa,t}|\mathcal{F}_{t-1} &\sim BP^{*}(\lambda_{\kappa,t,1}, \lambda_{\kappa,t,2}, \phi_{\kappa}), \\ \lambda_{\kappa,t} = (\lambda_{\kappa,t,1}, \lambda_{\kappa,t,2})^{T} &= \omega_{\kappa} + A_{\kappa}\lambda_{\kappa,t-1} + B_{\kappa}Y_{\kappa,t-1}. \end{aligned} \tag{3}$$

$$\begin{aligned} Y_{\kappa,t}|\mathcal{F}_{t-1} &\sim BP(\lambda_{\kappa,t,1}, \lambda_{\kappa,t,2}, \delta_{\kappa}), \\ \lambda_{\kappa,t} = (\lambda_{\kappa,t,1}, \lambda_{\kappa,t,2})^{T} &= \omega_{\kappa} + A_{\kappa}\lambda_{\kappa,t-1} + B_{\kappa}Y_{\kappa,t-1}. \end{aligned} \tag{4}$$

with $\phi_k \in R$, $\delta_{\kappa} \in R$, $\omega_{\kappa} = (\omega_{\kappa,1}, \omega_{\kappa,2})^{T}$, $A_{\kappa} = \begin{pmatrix} \alpha_{\kappa,11} & \alpha_{\kappa,12} \\ \alpha_{\kappa,21} & \alpha_{\kappa,22} \end{pmatrix}$, and $B_{\kappa} = \begin{pmatrix} \beta_{\kappa,11} & \beta_{\kappa,12} \\ \beta_{\kappa,21} & \beta_{\kappa,22} \end{pmatrix}$.

The regime switching is determined by:

$$S(t) = \begin{cases} 1 & \text{if } \|Z_{t-1}\|_{\infty} \leq c_1 \\ 2 & \text{if } c_1 < \|Z_{t-1}\|_{\infty} \leq c_2 \\ \vdots & \\ K-1 & \text{if } c_{K-2} < \|Z_{t-1}\|_{\infty} \leq c_{K-1} \\ K & \text{if } \|Z_{t-1}\|_{\infty} > c_{K-1} \end{cases} \tag{5}$$

where c_i (with $i \in 1, \ldots, K-1$) define the thresholds for regime changes, and $\|Z_{t-1}\|_{\infty} = \max\{Z_{t-1,1}, Z_{t-1,2}\}$.

We will later address the specific case where $K = 2$, which corresponds to the $SET - BPINGARCH\,(2,1,1)$ model with:

$$S(t) = \begin{cases} 1 & \text{if } \|Z_{t-1}\|_{\infty} \leq c_1 \\ 2 & \text{if } \|Z_{t-1}\|_{\infty} > c_1 \end{cases} \tag{6}$$

Remark 2.1**.** We can choose a different vector norm, such as $\|\cdot\|_1$, or select one of the components of Z_{t-1} to dictate the regime change.
Following the detailed model formulations, this work will provide a comprehensive analysis of the proposed models. This analysis will contain the development of estimation procedures for the $SET - BPINGARCH$ model. The statistical properties of these estimators will be assessed via simulation studies. Furthermore, the empirical relevance of the proposed model will be illustrated through an application on a real bivariate time series dataset.

Bibliographie

Armillotta, M., Fokianos, K. Nonlinear network autoregression. The Annals of Statistics 51, 2526-2552 (2023).

Bentarzi, M. and Sadoun, M. (2021). Efficient estimation in semiparametric self-exciting threshold $INAR$ processes. *Communications in Statistics-Simulation and Computation.* DOI: 10.1080/03610918.2021.1910300.

Cui, Y., Zhu, F. A new bivariate integer-valued GARCH model allowing for negative cross-correlation. TEST 27, 428-452 (2018).

Cui, Y., Li, Q., Zhu, F. Flexible bivariate Poisson integer-valued GARCH model. Annals of the Institute of Statistical Mathematics 72, 1449-1477 (2020).

Yang, K., Wang, D., Jia, B., & Li, H. (2018). An integer-valued threshold autoregressive process based on negative binomial thinning. Statistical Papers,59 (3), 1131 − 1160.

Karlis, D. Models for multivariate count time series. In: Handbook of discrete-valued time series, pp. 424 (CRC, 2016).

Ferland, R., Latour, A. and Oraichi, D. (2006). Integer-Valued $GARCH$ Process. *J. Time Ser. Anal.*, Vol. 27. No. 6, 923 − 942.

Fokianos, K., Støve, B., Tjøstheim, D., Doukhan, P. Multivariate count autoregression (2020).

Fokianos, K., Fried, R., Kharin, Y., Voloshko, V. Statistical analysis of multivariate discrete-valued time series. Journal of multivariate analysis 188, 104805 (2022).

Fokianos, K. Multivariate count time series modeling. Econometrics and Statistics (2021).

Liu, M., Zhu, F., Li, J., Sun, C. A systematic review of INGARCH models for integer-valued time series. Entropy 25, 922 (2023).

Liu, H. (2012). Some models for time series of counts. Columbia University.

Lee, S., Kim, D., Kim, B. Modeling and inference for multivariate time series of counts based on the INGARCH scheme. Computational Statistics and Data Analysis 177, 107579 (2023).

Session organisée par la Société Francophone de Classification

Une Approche Topologique de l'Analyse discriminante

Rafik Abdesselam

Laboratoires ERIC-COACTIS, Université Lumière Lyon 2, France
rafik.abdesselam@univ-lyon2.fr

Résumé. L'objectif de ce travail est de proposer une analyse discriminante topologique selon les différents types de données avec des variables explicatives quantitatives, qualitatives ou mixtes. Cette analyse topologique décisionnelle est une méthode de classification supervisée qui tente de découvrir les structures intrinsèques et les informations discriminantes intégrées dans les données. Il existe de nombreuses techniques prédictives, elles sont le plus souvent et le plus utilement appliquées à divers problèmes dans de nombreux domaines. L'approche topologique de discrimination proposée est basée sur la notion de graphes de voisinage dans un contexte décisionnel. Les variables explicatives sont plus ou moins corrélées ou liées selon que les variables sont de type quantitatives, qualitatives ou un mélange des deux. Ce modèle de discrimination topologique analyse la structure des corrélations ou dépendances observées dans chaque classe en fonction des variables explicatives. Pour valider l'efficacité de notre approche topologique, une série d'expériences est réalisée sur plusieurs jeux de données de référence UCI, avec des variables explicatives quantitatives, qualitatives et mixtes. Les résultats sont comparés à ceux des modèles d'analyse discriminante et de régression logistique.

Mots-clés. Mesure de proximité, graphe de voisinage, matrice d'adjacence, analyse discriminante.

Abstract. The objective of this work is to propose a topological discriminant analysis according to the type of data, quantitative, qualitative or mixed explanatory variables. This decisional topological classification is a supervised clustering method attempts to discover the intrinsic structures and discriminant information embedded in the data. There are many predictive techniques, they are most often and most usefully applied to various problems in many fields. The proposed topological approach of discrimination is based on the notion of neighborhood graphs in a decisional context. The explanatory variables are more or less correlated or linked depending on whether the variables type, quantitative, qualitative or a mixture of both. This topological model of discrimination analyzes the structure of the correlations or dependencies observed in each class according to the explanatory variables. To validate the effectiveness of our topological approach, a series of experiments are performed on several UCI benchmark datasets, with quantitative, qualitative and mixed explanatory variables. The results are compared to those of the discriminant analysis and logistic regression models.

Keywords. Proximity measure, neighborhood graph, adjacency matrix, discriminant analysis.

1 Introduction

L'analyse discriminante topologique (ADT) proposée est comparée à l'analyse discriminante (AD) qui fut pendant longtemps la méthode de classement utilisée dans de nombreux contextes, et à sa rivale, la régression logistique (LR) introduite plus récemment. L'AD est une méthode à la fois explicative et prédictive qui permet de vérifier si les groupes auxquels appartiennent les individus sont distincts, d'identifier quelles sont les caractéristiques des groupes sur la base de variables explicatives, de prédire le groupe d'appartenance d'un nouvel individu. Les applications de l'AD sont très nombreuses, allant de l'écologie et la médecine à la prédiction des risques en finance (credit scoring). L'AD suppose que les variables explicatives sont normalement distribuées et que les matrices de covariance intra-groupes sont égales. Cependant, elle est étonnamment résistante aux violations de ces hypothèses et constitue généralement un bon modèle de classification supervisée et de prise de décision. Il existe des approches spécifiques à l'AD, mais à notre connaissance aucune de ces approches n'a été proposée dans un contexte topologique. L'objectif de cet article est de proposer une approche topologique d'analyse discriminante (ADT) sur variables explicatives quantitatives, qualitatives ou mixtes. Le choix de la mesure de proximité parmi les nombreuses mesures existantes, joue un rôle important en analyse de données multidimensionnelles [Batagelj et Bren (1995), Zighed et al. (2012)]. Il a un fort impact sur les résultats de toute opération de structuration, de regroupement ou de classification d'objets. La structure de corrélation ou de dépendance des variables quantitatives ou qualitatives dépend des données considérées. Les résultats peuvent changer en fonction de la mesure de proximité choisie.

2 Contextes topologique de discrimination

L'analyse discriminante topologique consiste à analyser simultanément chaque table de données associée à chacune des modalités-classes de la variable cible à discriminer. On utilisera les notations suivantes :

- $Y_{(n,q)}$ est la matrice des données associée aux q variables binaires $\{y^k; k = 1, q\}$ de la variable qualitative cible à expliquer y à q modalités-groupes à discriminer,
- $X_{(n,p)}$ est la matrice des données associée aux p variables continues explicatives, ensemble de p variables discriminantes $\{x^j; j = 1, p\}$, à $n = \sum_{k=1}^{q} n_k$ lignes-individus et p colonnes-variables,
- $X_{k(n_k,p)}$ est la matrice des données associée aux p variables explicatives à n_k individus ayant la $k^{ième}$ modalité de y,
- $Z_{(n,r)}$ est la matrice des données associée aux r variables qualitatives explicatives, ensemble de r variables discriminants $\{z^j; j = 1, r\}$, à $n = \sum_{k=1}^{r} n_k$ lignes-individus et r colonnes–variables qualitatives,
- $Z_{r(n_k,r)}$ est la matrice des données associée aux r variables qualitatives à n_k individus ayant la $k^{ième}$ modalité de y.

Etant donnée une mesure de proximité u_k, on peut définir une relation de voisinage, V_{u_k}, comme étant une relation binaire basée sur $E_k \times E_k$, $E_k = \{x^1, \cdots, x^j, \cdots, x^p\}$ étant l'ensemble des p variables explicatives. Plusieurs définitions sont possibles pour construire

cette relation de voisinage binaire, on peut choisir l'Arbre de Longueur Minimal(ALM)[Kim et Lee (2003)], le Graphe de Gabriel (GG)[Park et al. (2006)] ou, comme c'est le cas ici, le Graphe de Voisinage Relatif (GVR)[Toussaint (1980)]. Étant donné un ensemble E_k de p variables du tableau de données X_k et une mesure de proximité u_k, pour des données continues ou binaires, on construit la matrice d'adjacence associée V_{u_k} d'ordre p, où toutes les paires de variables voisines dans E_k satisfont la propriété GVR suivante :

$$V_{u_k}(x^l\ ,\ x^r) = \begin{cases} 1 & \text{if } u_k(x^l\ ,\ x^r) \ \leq \ \max[u_k(x^l\ ,\ x^t), u_k(x^t\ ,\ x^r)]\ ; \\ & \qquad\qquad \forall x^l,\ x^r,\ x^t \in E,\ x^t \neq x^l \ \ and \ \ x^t \neq x^r \\ 0 & \text{otherwise.} \end{cases}$$

Cela signifie que si deux variables x^l et x^r qui vérifient la propriété GVR sont connectées par une arête, les sommets x^l et x^r sont voisins.

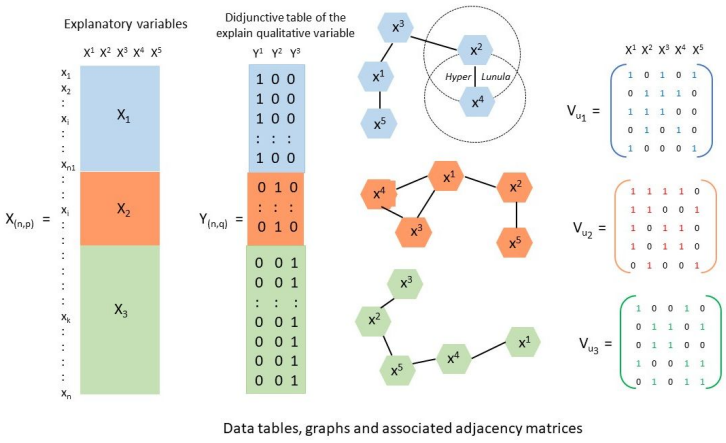


Figure 1: GVR - Matrices d'adjacence pour la discrimination topologique

La Figure 1 présente un exemple illustratif avec cinq variables explicatives quantitatives $\{x^j; j = 1, 5\}$ et $y = \{y^k; k = 1, 3\}$ une variable qualitative cible à expliquer à trois modalités-groupes. A partir des tableaux de données de chaque groupe $X_1 = (X/Y = 1), X_2 = (X/Y = 2)$ et $X_3 = (X/Y = 3)$, on établit les matrices d'adjacence binaires associées V_{u_1}, V_{u_2} et V_{u_3}, selon la structure de voisinage et les mesures de proximité choisies u_1, u_2 et u_3.
Par exemple, pour le tableau de données X_1, on voit que pour la première et la quatrième variable, $V_{u_1}(x^1, x^4) = 1$, cela signifie que sur le plan géométrique, l'hyper-Lunule (intersection entre les deux hypersphères centrées sur les deux variables x^1 et x^4) est vide.

2.1 Matrices d'adjacence de référence

L'objectif est d'analyser de manière topologique les structures de corrélation ou de dépendance des variables explicatives des données considérées. Les expressions des matrices de référence d'adjacence appropriées sont décrites dans [Abdesselam (2021)] en fonction du type de variables explicatives considérées, quantitatives ou qualitatives ou un mélange des deux.

2.1.1 Variables explicatives quantitatives

Nous construisons la matrice d'adjacence notée $V_{u^\star}$, qui correspond le mieux à la matrice de corrélation. Ainsi, pour examiner la structure de corrélation entre les variables, nous regardons la significativité de leur coefficient de corrélation linéaire. Cette matrice d'adjacence peut s'écrire comme suit en utilisant le test t ou le test t de Student du coefficient de corrélation linéaire de Bravais-Pearson :

Pour chaque tableau de données $X_k = (X/Y = k)$, nous construisons la matrice d'adjacence de référence notée $V_{u_k^\star}$ à partir de la matrice de corrélation du tableau de données X_k.

Pour chaque tableau de données X_k, la matrice d'adjacence de référence $V_{u_k^\star}$ associée à la mesure de référence $u_k^\star$ est définie ainsi :

$$V_{u_k^\star}(x^l, x^r) = \begin{cases} 1 & \text{if p-value } = P[\mid T_{n-2} \mid > \text{t-value }] \leq \alpha \ ; \ \forall l, r = 1, p \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Où, T_{n-2} désigne la loi de Student à $\nu = n - 2$ degrés de liberté et la p-value, le seuil de signification du test du coefficient de corrélation linéaire.

La matrice d'adjacence de référence $V_{u_k^\star}$ ainsi construite est associée à une mesure de proximité de référence inconnue, notée $u_k^\star$.

2.1.2 Variables qualitatives explicatives

L'objectif est d'analyser la structure de dépendance à savoir s'il existe une association topologique entre toutes ces variables qualitatives. Pour chaque tableau de données Z_k, on construit la matrice d'adjacence de référence $V_{u_k^\star}$, selon l'association significative entre les variables du tableau Z_k, obtenue à partir du tableau de Burt $\mathcal{B}_k$.

Pour chaque table de données Z_k, la matrice d'adjacence de référence $V_{u_k^\star}$ associée à la mesure de référence $u_k^\star$ est ainsi définie :

$$V_{u_k^\star}(z^{ht}, z^{ls}) = \begin{cases} 1 & \text{if } \frac{\mathcal{B}_{ht\,ls}}{\mathcal{B}_{ht\,..}} \geq \frac{\mathcal{B}_{ht\,..}}{nq^2}; \ \ \forall h, l = 1, r \ ; \ t = 1, m_h \ and \ s = 1, m_l \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Ainsi, pour examiner les similitudes entre les modalités des variables, nous examinons l'écart entre chaque profil-modalité et son profil moyen, c'est-à-dire l'écart vers l'indépendance.

2.1.3 Variables explicatives mixtes

Le traitement simultané de données mixtes (quantitatives et qualitatives) ne peut être réalisé directement par les méthodes conventionnelles d'analyse de données. Ainsi, nous transformons tout d'abord les données qualitatives en données quantitatives [Abdesselam (2008)]. Cette transformation est basée sur l'analyse de variance multivariée (MANOVA) et sur la maximisation du critère mixte, proposé en termes de carrés de corrélation par Tenenhaus [Tenenhaus (1977)] et géométriquement en termes de cosinus carrés des angles par Escofier [Escofier et Pagès (1985)].

Dans un deuxième temps, on construit la matrice d'adjacence $V_{u^\star}$, associée à la mesure de proximité de référence $u^\star$, à partir de la matrice de corrélation de toutes les variables, quantitatives et qualitatives transformées, selon l'expression (1).

2.2 Analyse Discriminante Topologique

On utilisera les notations matricielles suivantes :

$$X_{(n,p)} = \begin{pmatrix} X_{1(n_1,p)} \\ \vdots \\ X_{k(n_k,p)} \\ \vdots \\ X_{q(n_q,p)} \end{pmatrix} \quad \mathrm{Z}_{(n,r)} = \begin{pmatrix} Z_{1(n_1,r)} \\ \vdots \\ Z_{k(n_k,r)} \\ \vdots \\ Z_{q(n_q,r)} \end{pmatrix} \quad \mathrm{Y}_{(n,q)} = \begin{pmatrix} Y_{1(n_1,q)} \\ \vdots \\ Y_{k(n_k,q)} \\ \vdots \\ Y_{q(n_q,q)} \end{pmatrix} \quad \mathrm{V}_{u^\star(qp,p)} = \begin{pmatrix} V_{u_1^\star(p)} \\ \vdots \\ V_{u_k^\star(p)} \\ \vdots \\ V_{u_q^\star(p)} \end{pmatrix} \quad \widehat{X}_{(n,p)} = \begin{pmatrix} X_1 V_{u_1^\star} \\ \vdots \\ X_k V_{u_k^\star} \\ \vdots \\ X_q V_{u_q^\star} \end{pmatrix}$$

- $V_{u_k^\star}$ est la matrice symétrique d'adjacence d'ordre p, associée à la mesure de proximité de référence $u_k^\star$, qui résume le mieux la structure des corrélations du tableau de données X_k ou la structure de dépendance du tableau de données Z_k,
- $\widehat{X}_{(n,p)} = Diag[X]V_{u^\star} = XV_{u^\star}$ est la matrice des données projetées à n individus et p variables,
- M_p est la matrice des distances d'ordre p dans l'espace des individus,
- $D_n = \frac{1}{n} I_n$ est la matrice diagonale des poids d'ordre n dans l'espace des variables.

On analyse d'abord de manière topologique, la structure de corrélation des variables à l'aide d'une ACP topologique, qui consiste à réaliser l'ACP [**?**, **?**] du triplet $(\widehat{X}, M_p, D_n)$ de la matrice de données projetée $\widehat{X} = XV_{u^\star}$, puis nous procédons à une analyse discriminante sur les composantes principales significatives de l'ACP topologique précédente.
L'approche ADT proposée, consiste à effectuer une analyse discriminante sur les facteurs significatifs de l'analyse en composantes principales topologiques du triplet $(\widehat{X}, M_p, D_n)$.

3 Exemple illustratif

Pour illustrer l'approche ADT, on a établi un Benchmark de plusieurs jeux de données extraits du référentiel UCI Machine Learning sur différents thèmes et avec différentes dimensions des données [UCI (1988), Govaert (2003)]. L'approche topologique proposée a été testée sur sept bases de données réelles, les résultats obtenus, présentés dans le Tableau 1, ont été comparés à ceux de l'analyse discriminante et de la régression logistique. L'exemple illustratif traite les caractéristiques bancaires des clients d'une agence bancaire (jeu de données n°2).

Les statistiques descriptives des variables explicatives mixtes sont présentées dans le tableau 2 ainsi que celles de la variable cible à discriminer dans le tableau 3.

Table 1: Benchmark - Bases de données

Explanatory Variables	Data	n	p	q	Well classified (%) TDA	DA	LR
1 - Continuous	Iris	150	4	3	88.00	98.00	**98.67**
2 - Mixed	Credit bank	468	8	2	**97.01**	88.46	76.28
3 - Continuous	Brands of bottled water	38	8	2	89.47	89.47	**94.74**
4 - Continuous	Wine	178	13	3	**100.00**	99.44	**100.00**
5 - Categorical	Bank customer	3808	6	2	**90.21**	62.61	89.15
6 - Continuous	Raisin	900	6	2	**100.00**	86.33	**100.00**
7 - Continuous	Wine Quality	6497	12	2	**99.65**	99.48	99.54

Table 2: Statistiques sommaires des variables bancaires explicatives mixtes

Credit bank Data Continuous variables	Mean	Standard Deviation	Coefficient of variation	Min	Max
Savings Amount	1040.79	2884.77	2.77	0.00	21000.00
Years Seniority	6.17	5.35	0.87	0.50	28.00
Average outstanding	759.18	381.24	0.50	145.00	2315.00
Average Banking Transactions	5277.01	3950.83	0.75	450.00	17500.00
Average cumulative debits	70.02	43.50	0.62	8.00	187.00

Modalities of Categorical variables	Frequency	Percentage	Cummuled
Domiciled salary	316	67.52	67.52
Non-domiciled salary	152	32,48	32.48
Total	468	100.00	100.00
Authorized-Overdraft	202	43.16	43.16
Prohibited-Overdraft	266	56.84	56.84
Total	468	100.00	100.00
Authorized-Checkbook	415	88.68	88.68
Prohibited-Checkbook	53	11.32	11.32
Total	468	100.00	100.00

Table 3: Statistiques de la variable cible

Modality	Frequency	Percentage	Cummuled
Good customers	237	50.64	50.64
Bad customers	231	49.36	49.36
Total	468	100.00	100.00

La matrice d'adjacence de référence globale $V_{u^\star}$ présentée dans le Tableau 4 est associée à la mesure de proximité $u^\star$ la plus adaptée aux données considérées, elle est construite à partir des matrices de corrélation des tableaux de données X_1 et X_2 selon l'expression (1). Dans ce cas de variables explicatives mixtes, les données qualitatives ont d'abord été transformées en données quantitatives, puis on a considéré toutes les données quantitatives, c'est-à-dire toutes les variables quantitatives et qualitatives transformées.
Notons que deux variables quantitatives corrélées positivement sont liées et deux variables corrélées négativement sont liées, mais distantes, nous prendrons donc en compte le signe de la corrélation entre les variables dans la matrice d'adjacence.
On a effectué une ACP Topologique non normée pour identifier la structure de corrélation de toutes les variables, une analyse discriminante est ensuite appliquée sur les composantes principales significatives de cette ACP Topologique des données projetées.
La figure 2 présente sur le premier plan factoriel de l'ACP Topologique, les corrélations entre les composantes principales-facteurs et les variables initiales. Les deux premiers facteurs de l'ACP topologique expliquent respectivement $79,84\%$ et $19,68\%$, soit $99,53\%$ de la variation totale de l'ensemble des données, ils fournissent une synthèse adéquate des données mixtes, c'est-à-dire des caractéristiques bancaires des clients de l'agence.

Table 4: Matrices globales de correlations et d'adjacence de référence

$$R = \left(\begin{array}{c|c} \mathrm{R}_1 & 0 \\ \hline 0 & \mathrm{R}_2 \end{array}\right) \qquad \mathrm{V}_{u^*} = \left(\begin{array}{c|c} \mathrm{V}_{u_1^*} & 0 \\ \hline 0 & \mathrm{V}_{u_2^*} \end{array}\right)$$

$$V_{u^*} = \left(\begin{array}{ccccccccccc|ccccccccccc} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & -1 & 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & -1 & 0 & 0 & 0 & -1 & 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & -1 & 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & -1 & 0 & 0 & 0 & -1 & 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \end{array}\right)$$

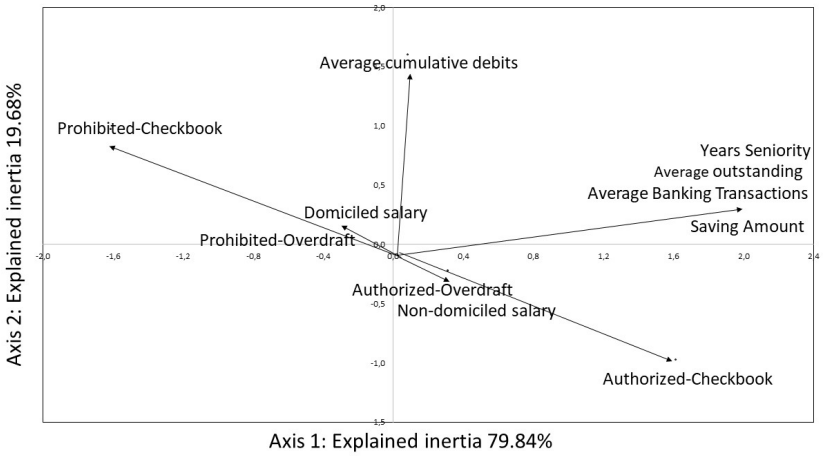


Figure 2: Représentation des variables explicatives mixtes

Le tableau 5 résume les profils significatifs des deux groupes de clients ; avec un risque d'erreur inférieur ou égal à 5%. Pour établir les profils des groupes de clients de l'agence bancaire, une Analyse Discriminante de Fischer a été appliquée sur les variables explicatives mixtes.

Les coefficients de la fonction discriminante qui sépare au mieux les deux groupes de clients, ont été classés, selon la valeur du test t-student, dans le Tableau 5.

Le premier groupe composé de 237 "bons clients" est caractérisé par une domiciliation en agence du salaire du client et des autorisations de découvert et de chéquier. Le second groupe constitué de 231 "mauvais clients" se caractérise par une combinaison de débits moyens élevés, de salaire non domicilié dans l'agence et d'interdictions de découvert et de chéquier.

Table 5: ADT - Topological Discriminant Analysis on mixed variables

Fisher linear function Variable	Coefficient function Discriminant	Standard Deviation	Ratio t-Student	Proba p-value
Good customers				
Domiciled salary	0.0000	0.0000	3.40**	0.00034
Authorized-Overdraft	0.0000	0.0000	3.37**	0.00038
Authorized-Checkbook	0.0001	0.0000	15.44**	0.00001
Bad customers				
Savings Amount	-0.0000	0.0000	-0.94	0.23860
Years Seniority	-0.0000	0.0000	-0.94	0.23860
Average outstanding	-0.0000	0.0000	-0.94	0.23860
Average Banking Transactions	-0.0000	0.0000	-0.94	0.23860
Average cumulative debits	-0.0001	0.0000	-19.76**	0.00001
Non-domiciled salary	-0.0000	0.0000	-3.40**	0.00034
Prohibited-Overdraft	-0.0000	0.0000	-3.37**	0.00038
Prohibited-Checkbook	-0.0001	0.0000	-15.44**	0.00001
Constant	-0.668013	-0.147380		
R2 = 0.5829	F = 324.8962	PROBA = 0.001		
D2 = 5.5666	T2 = 651.1898	PROBA = 0.001		

Significance level α : $^{**}\alpha \leq 1\%$; $^{*}\alpha \in]1\%; 5\%]$

Table 6: Matrices de confusion & Courbes ROC

TDA

Actual classification	Predicted classification: Good customers	Bad customer	Total
Good customers	228	9	237
Bad customers	5	226	231
Total	233	235	468

Pourcentage de bien classés : 97.01% AUC - ROC : 0.974

ADM

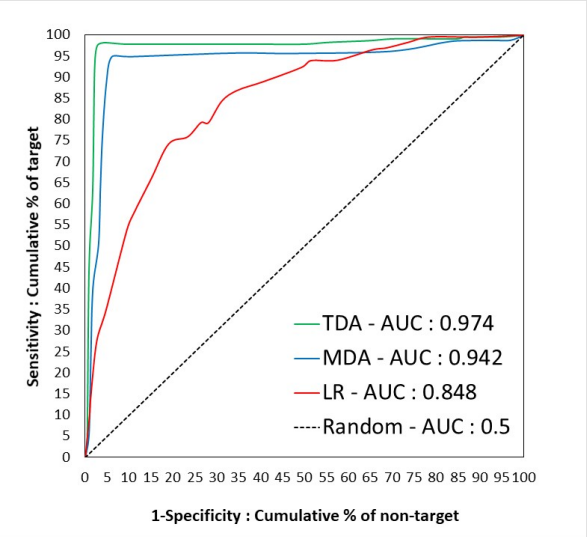
Actual classification	Predicted classification: Good customers	Bad customer	Total
Good customers	192	45	237
Bad customers	9	222	231
Total	201	267	468

Pourcentage de bien classés : 88.46% AUC - ROC : 0.942

RL

Actual classification	Predicted classification: Good customers	Bad customers	Total
Good customers	187	50	237
Bad customers	61	170	231
Total	248	220	468

Pourcentage de bien classés : 76.28% AUC - ROC : 0.848



À titre de comparaison, le tableau 6 présente les résultats des matrices de confusion du modèle d'Analyse Discriminante Mixte (ADM) [**?**] et de régression logistique (LR). Les courbes ROC (Receiver Operating Characteristic) et les scores ROC-AUC illustrent les performances des modèles.

4 Conclusion

Cet article propose une approche topologique de l'analyse discriminante (ADT) qui peut enrichir les méthodes classiques d'analyse de données dans le cadre des modèles prédictifs. Les résultats de l'ADT, basée sur la notion de graphe de voisinage, obtenus à partir d'un

benchmark de base de données, sont aussi bons, voire meilleurs, selon le critère du pourcentage de bien classés, que ceux de l'ADM et de la RL. Il serait intéressant de les comparer aux autres méthodes de classement existantes. La méthode de classement ADT peut être mise en œuvre à partir des procédures d'analyse en composantes principales et de discrimination des logiciels SAS, SPAD ou R. Il serait intéressant d'étendre cette approche topologique à d'autres modèles prédictifs d'analyse de données, notamment dans le cadre de la régression multiple.

Bibliographie

Abdesselam, R. (2021) *A Topological Clustering of variables.* Journal of Mathematics and System Science. David Publishing Company,Vol.11, Issue 2, pp.1-17, 2021.

Abdesselam, R. (2008) *Analyse en Composantes Principales Mixte.* Classification : points de vue croisés, RNTI-C-2, Cépaduès Editions, 31-41, 2008.

Batagelj, V., Bren, M. (1995) *Comparing resemblance measures.* In Journal of classification, 12, 73–90, 1995.

Escofier, B. et Pagès, J. (1985) Mise en oeuvre de l'AFM pour des tableaux numériques, qualitatifs, ou mixtes. Publication interne de l'IRISA, 429, 1985.

Govaert, G.: Analyse des données (2003). Hermes Science, Lavoisier.

Kim, J.H. and Lee, S. (2003) Tail bound for the minimal spanning tree of a complete graph. *In Statistics & Probability Letters*, 4, 64, 425–430.

Lebart, L. (1989) *Stratégies du traitement des données d'enquêtes.* La Revue de MODULAD, 3, 21–29, 1989.

Panagopoulos, D. (2022) *Topological data analysis and clustering.* Chapter for a book, Algebraic Topology (math.AT) arXiv:2201.09054, Machine Learning.

Park, J. C., Shin, H. and Choi, B. K. (2006) *Elliptic Gabriel graph for finding neighbors in a point set and its application to normal vector estimation.* In Computer-Aided Design Elsevier, 38, 6, 619–626.

Tenenhaus, M. (1977) Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de statistique appliquée*, tome 25, no 2, 39–56, 1977.

Toussaint, G. T. (1980) *The relative neighbourhood graph of a finite planar set.* In Pattern recognition, 12, 4, 261–268, 1980.

UCI Machine Learning Repository (1988), https://archive.ics.uci.edu/datasets.

Ward, J.R. (1963) *Hierarchical grouping to optimize an objective function.* In Journal of the American statistical association JSTOR, 58, 301, 236–244, 1963.

Zighed, D., Abdesselam, R., and Hadgu, A. (2012) *Topological comparisons of proximity measures.* In the 16th PAKDD 2012 Conference. In P.-N. Tan et al., Eds. Part I, LNAI 7301, Springer-Verlag Berlin Heidelberg, 379–391, 2012.

Sur les structures de clustering multi-niveau définies par des convexités d'intervalle. Applications à la hiérarchie du lien simple

Patrice Bertrand [1] & Jean Diatta [2]

[1] *Ceremade, Université Paris Dauphine - PSL, France*
patrice.bertrand@ceremade.dauphine.fr
[2] *LIM-EA2525, Université de La Réunion, France*
jean.diatta@univ-reunion.fr

Résumé. La classification hiérarchique, et la plupart des structures de clustering multi-niveau, peuvent être caractérisées, dans un cadre unifiant, comme étant des convexités induites par des fonctions d'intervalle vérifiant des propriétés spécifiques. Dans ce qui suit, ce cadre théorique unificateur se révèle être approprié à la pratique de l'analyse des données. Tout d'abord, nous proposons une caractérisation flexible des hiérarchies et des hiérarchies faibles en tant que convexités d'intervalle d'un certain type, et en utilisant cette approche, nous caractérisons les structures particulières que sont les hiérarchies d'Apresjan et les hiérarchies faibles de Bandelt et Dress. Par ailleurs, nous introduisons une séquence de dissimilarités définies à partir de chemins, que l'on peut calculer récursivement, et qui est décroissante d'une dissimilarité arbitraire jusqu'à son ultramétrique sous-dominante. Les dissimilarités définies ainsi à partir de chemins valués définissent deux séquences de familles imbriquées de convexités d'intervalle. L'une des deux séquences est croissante de la hiérarchie d'Apresjan à la hiérarchie du lien simple. Une application à la simplification et à la validation de la hiérarchie du lien simple sera présentée.

Mots-clés. Hiérarchie d'Apresjan, convexité d'intervalle, hiérarchie du lien simple, hiérarchie faible, dissimilarité définie à partir de chemins

Abstract. Hierarchies of sets, and most multilevel clustering structures, have been characterized as convexities induced by interval functions satisfying specific properties, thus giving rise to a unifying framework for characterizing multilevel clustering structures. Here, we show that this unifying framework can be relevant to data mining practice. First, we provide a flexible characterization of hierarchies and weak hierarchies as interval convexities, and using this approach, we characterise the particular structures that are Apresjan hierarchies and Bandelt and Dress weak hierarchies. Furthermore, we introduce a sequence of path-based dissimilarities, which can be computed recursively, and which is decreasing from an arbitrary dissimilarity to its subdominant ultrametric. These path-based dissimilarities define two sequences of nested families of interval convexities. One of the two sequences grows from the Apresjan hierarchy to the single-link hierarchy. An application to the simplification and validation of the single-link hierarchy will be presented.

Keywords. Apresjan hierarchy, interval convexity, single-link hierarchy, weak hierarchy, path-based dissimilarities

1 Introduction

Le clustering, appelé aussi classification non supervisée, est une technique d'exploration des données à la fois populaire et utile qui vise à révéler une structure cachée au sein d'un ensemble de données. La structure révélée est généralement exprimée sous la forme d'une collection de sous-ensembles homogènes appelés *classes*. Selon Mirkin et Muchnick (1998), une des principales approches en clustering, consiste à déterminer les classes directement par leur définition. Ce texte s'intéresse aux classes qui sont définies comme étant les parties convexes au sens d'une *fonction d'intervalle*. Une fonction d'intervalle I sur un ensemble de données S est une fonction symétrique qui à chaque couple $(x, y) \in S^2$ associe un sous-ensemble $I(x, y)$ de S, contenant x et y. Le sous-ensemble $I(x, y)$ est appelé *I-intervalle d'extrémités x et y*. Un sous-ensemble A de S est dit *convexe au sens de la fonction d'intervalle I*, ou simplement *I-convexe*, s'il contient chaque I-intervalle dont les extrémités appartiennent à A. La collection $\mathscr{C}_I$ des sous-ensembles I-convexes de S forme une convexité (au sens abstrait, e.g. Van de Vel (1993)) qui est appelée *convexité d'intervalle induite par I* (cf. Calder (1971)). Dans ce texte, nous introduisons une séquence de dissimilarités définies à partir de chemins, qui est décroissante d'une dissimilarité arbitraire jusqu'à son ultramétrique sous-dominante. De cette suite décroissante de dissimilarités, nous déduisons une suite de convexités qui croit (au sens de l'inclusion) de la hiérarchie d'Apresjan à la hiérarchie du lien simple. Nous en déduisons une procédure pour simplifier un dendrogramme représentant la hiérarchie du lien simple, en supprimant les classes qui sont les moins pertinentes au sens d'un critère de même nature que celui du lien simple.

2 Préliminaires

Soit $\mathscr{C}$ une collection de sous-ensembles d'un ensemble fini S. La collection $\mathscr{C}$ est appelée *classification multi-niveau* si $\mathscr{C}$ contient S mais pas $\emptyset$, et si $\mid \mathscr{C} \mid \geq 2$. La collection $\mathscr{C}$ est dite *hiérarchique* si pour tout $A, B \in \mathscr{C}, A \cap B \in \{A, B, \emptyset\}$. Par ailleurs, $\mathscr{C}$ est dite *faiblement hiérarchique* si pour tout $A, B, C \in C, A \cap B \cap C \in \{A \cap B, A \cap C, B \cap C\}$. Enfin, on dit que $\mathscr{C}$ est *une convexité* si $\mathscr{C}$ contient $\{\emptyset, S\}$ et est fermée par intersections.

Une application $g : S^2 \mapsto 2^S$ est appelée *fonction d'intervalle* si pour tout $x, y \in S$, on a $\{x, y\} \subseteq g(x, y) = g(y, x)$. Si g est une fonction d'intervalle, alors une partie C de S est dite g-convexe, si $g(x, y) \subseteq C$ pour tout $x, y \in C$. L'ensemble des sous-ensembles g-convexes est une convexité, dite *convexité d'intervalle*, notée par la suite $\mathrm{conv}(g)$. On dira que $\mathrm{conv}(g)$ est la *convexité induite* par g.

Soit une application $g : S^2 \mapsto 2^S$ telle que $\{x, y\} \subseteq g(x, y)$, pour tout $x, y \in S$. Par la suite, nous notons (H), (W) et (W$'$) les conditions définies par :

(H) Pour tout $x_1, x_2, x_3 \in S$, $g(x_1, x_2) \subseteq g(x_1, x_3)$ ou $g(x_1, x_3) \subseteq g(x_1, x_2)$

(W) Pour tout $x_1, x_2, x_3 \in S$, il existe i, j, k avec $\{i, j, k\} = \{1, 2, 3\}$ tels que :
$$g(x_i, x_j) \subseteq g(x_i, x_k) \text{ et } g(x_k, x_j) \subseteq g(x_k, x_i).$$

(W′) Pour tout $x_1, x_2, x_3 \in S$, il existe i, j, k avec $\{i, j, k\} = \{1, 2, 3\}$, tels que :
$$g(x_i, x_j) \subseteq g(x_i, x_k),\ g(x_k, x_j) \subseteq g(x_k, x_i) \text{ et } g(x_j, x_i) \subseteq g(x_j, x_k).$$

On vérifie que (H) n'implique pas (W) et (W) n'implique pas (W′). De plus :

$$(\mathrm{W}') \Leftrightarrow (\mathrm{H}) \wedge (\mathrm{W}).$$

Par ailleurs, (H) implique (W) si g est symétrique. Par la suite, nous notons J_g et M_g les applications définies par

$$J_g = g \cup \overline{g} \text{ et } M_g = g \cap \overline{g},$$

où $\overline{g}$ est l'application définie par $\overline{g}(x, y) = g(y, x)$ pour tout $x, y \in S$.

La caractérisation suivante des classifications hiérarchiques et faiblement hiérarchiques, est flexible dans le sens où elle ne nécessite pas que g soit une fonction d'intervalle.

Proposition 2.1 *Soit $\mathscr{C}$ une classification multi-niveau sur S.*

(i) *$\mathscr{C}$ est hiérarchique si et seulement s'il existe une application g qui vérifie* (H) *et telle que $\mathscr{C} = \mathrm{conv}(J_g)$.*

(ii) *$\mathscr{C}$ est faiblement hiérarchique si et seulement s'il existe une application g qui vérifie* (W) *et telle que $\mathscr{C} = \mathrm{conv}(M_g)$.*

Par la suite, δ désigne une dissimilarité définie sur S, i.e. δ est une application de S^2 dans $\mathbb{R}^+$ telle que $\delta(x, y) = \delta(y, x) \geq \delta(x, x) = 0$ pour tout $x, y \in S$. Une partie C de S est appelée *classe d'Apresjan* (cf. Apresjan (1966), Benzecri (1972), Diatta and Fichet (1998)), si pour tout $x, y \in C$, on a $\delta(x, y) < \min\limits_{z \notin C} \big(\min\{\delta(x, z), \delta(y, z)\}\big)$.
Une partie C de S est appelée *classe faible* (cf. Bandelt and Dress (1989), Diatta and Fichet (1998)) si pour tout $x, y \in C$, on a $\delta(x, y) < \min\limits_{z \notin C} \big(\max\{\delta(x, z), \delta(y, z)\}\big)$.
L'ensemble des classes d'Apresjan d'une dissimilarité δ forme une hiérarchie sur S. Cette hiérarchie, notée $\mathcal{H}_A(\delta)$, est appelée *hiérarchie d'Apresjan*. L'ensemble des classes faibles d'une dissimilarité δ forme une hiérarchie faible sur S. Cette hiérarchie faible, notée $\mathcal{W}_{\mathcal{BD}}(\delta)$, est appelée *hiérarchie faible de Bandelt and Dress*.
Par la suite, $g_\delta(x, y)$ désigne la boule fermée de centre x et de rayon $\delta(x, y)$, i.e.

$$g_\delta(x, y) := B_\delta(x, \delta(x, y)) = \{s \in S \mid \delta(x, s) \leq \delta(x, y)\}.$$

En posant $J_\delta := J_{g_\delta}$ et $M_\delta := M_{g_\delta}$, on vérifie que $\mathcal{H}_\mathcal{A}(\delta) = \mathrm{conv}(J_\delta) \subseteq \mathcal{W}_{\mathcal{BD}}(\delta) = \mathrm{conv}(M_\delta)$, donc la hiérarchie d'Apresjan est incluse dans la hiérarchie faible de Bandelt and Dress.

3 Une filtration de la hiérarchie du lien simple

Soit $P = u_0 u_1 \dots u_m$ un chemin simple du graphe complet K_S défini sur S et valué par δ. On note $\mathrm{val}_\delta(P)$ le saut maximum au sens de δ le long du chemin P entre u_0 et u_m. Formellement,

$$\mathrm{val}_\delta P := \begin{cases} \max\limits_{1 \leq i \leq m} \delta(u_{i-1}, u_i), \text{ si } u_0 \neq u_m, \\ 0, \text{ sinon, i.e. si } P \text{ est réduit à un sommet } u_0 = u_m. \end{cases}$$

Pour tout $x, y \in S$ et tout $\ell \geq 1$, on note $\mathcal{P}_{x-y}^{(\ell)}$ l'ensemble des chemins (simples et sans boucles) de longueur au plus ℓ et qui relient x à y. De plus, on note $\delta_\ell(x,y) := \min\{\mathrm{val}_\delta(P) \mid P \in \mathcal{P}_{x-y}^{(\ell)}\}$. Xu *et al.* (2007) et Yu *et al.* (2014, 2016) ont utilisé cette dissimilarité δ_ℓ et l'ont appelée *distance transitive d'ordre au plus* ℓ.

Théorème 3.1 *Soit ℓ_1, ℓ_2 des entiers strictement strictement positifs, $\ell \in \{2, \ldots, n-1\}$ et δ une dissimilarité arbitraire définie sur S.*

(i) *Si $\ell_1 \leq \ell_2$, alors $\delta_{\ell_1} \succeq \delta_{\ell_2}$ où $\succeq$ désigne l'ordre entre dissimilarités défini points par points, i.e. $\delta_{\ell_1} \succeq \delta_{\ell_2}$ si $\delta_{\ell_1}(x,y) \geq \delta_{\ell_2}(x,y)$ pour tout $x, y \in S$.*

(ii) *Si $\ell_1, \ell_2 \geq n-1$, alors $\delta_{\ell_1} = \delta_{\ell_2} = \delta_{n-1}$ qui est la sous-dominante ultramétrique de δ.*

(iii) $\delta_\ell(x,y) = \min\limits_{z \in S} \max\{\delta_{\ell-1}(x,z), \delta(z,y)\}$, *pour tout $x, y \in S$.*

(iv) *Si $\delta_{\ell-1} = \delta_\ell$, alors $\delta_{\ell-1} = \delta_\ell = \delta_{\ell+1} = \ldots = \delta_{n-1}$.*

On pose $r(\delta) = \min\{\ell \in [1, n-1] \mid \delta_\ell = \delta_{\ell+1}\}$, et par la suite, on adopte les notations simplifiées suivantes : $J_\ell = J_{\delta_\ell}$ et $M_\ell = M_{\delta_\ell}$.

Théorème 3.2 *Pour toute dissimilarité δ sur S, on a :*

(i) *La convexité induite par J_{n-1} est la hiérarchie du lien simple de δ.*

(ii) *Soi $\ell_1, \ell_2 \in \{1, \ldots, n-1\}$. Si $\ell_1 \leq \ell_2$, alors* $\mathrm{conv}(J_{\ell_1}) \subseteq \mathrm{conv}(J_{\ell_2})$.

Il est facile de vérifier que si I_1 et I_2 sont des fonctions d'intervalles, alorsi $I_1 \subseteq I_2$ implique $\mathrm{conv}(I_2) \subseteq \mathrm{conv}(I_1)$. Néanmoins, $J_{\ell_2} \not\subseteq J_{\ell_1}$ en général, et la propriété (ii) ci-dessus est non triviale : voir Bertrand et Diatta (2024) pour une preuve. Par la suite, la hiérarchie du lien simple de δ est notée $\mathcal{H}_{\mathrm{SL}}(\delta)$, et $\mathcal{H}_\ell(\delta)$ désigne la hiérarchie d'Apresjan de δ_ℓ. Ainsi, on a :

$$\mathcal{H}_\ell(\delta) = \mathrm{conv}(J_\ell) \; ; \; \mathcal{H}_{\mathcal{A}}(\delta) = \mathcal{H}_1(\delta) \; ; \; \mathcal{H}_{\mathrm{SL}}(\delta) = \mathcal{H}_{n-1}(\delta) = \mathrm{conv}(J_{n-1}).$$

Par ailleurs, pour toute classe $C \in \mathcal{H}_{\mathrm{SL}}(\delta)$, on note $\lambda(C,\delta) := \min\{\ell \,|\, C \in \mathcal{H}_\ell(\delta)\}$. L'entier positif $\lambda(C,\delta)$, noté aussi simplement $\lambda(C)$, est appelé *longueur minimale de chemin* de C. La plus grande des longueurs minimales de chemin de classes appartenant à la hiérarchie du lien simple est notée $\overline{\lambda}(\delta)$, i.e. $\overline{\lambda}(\delta) := \max\limits_{X \in \mathcal{H}_{\mathrm{SL}}(\delta)} \lambda(X,\delta)$.

Corollaire 3.3 *If C est une classe de la hiérarchie du lien simple de δ, alors*

(i) $1 \leq \lambda(C) \leq \overline{\lambda} \leq r(\delta) \leq n-1$.

(ii) $\mathcal{H}_{\mathcal{A}}(\delta) = \mathcal{H}_1(\delta) \subseteq \ldots \subseteq \mathcal{H}_{\overline{\lambda}-1}(\delta) \subset \mathcal{H}_{\overline{\lambda}}(\delta) = \ldots = \mathcal{H}_{n-1}(\delta) = \mathcal{H}_{\mathrm{SL}}(\delta)$.

4 Applications

Dans cette section, nous montrerons comment les résultats précédents permettent d'une part de définir une procédure de simplification du dendrogramme du lien simple, et, d'autre part, d'ordonner les classes de la hiérarchie du lien simple selon un critère de qualité de même nature que celui du lien simple.

On rappelle que $\delta = \delta_1 \succ \ldots \succ \delta_\ell \succ \ldots \succ \delta_{r(\delta)} = \ldots = \delta_{n-1}$. Soit $\Delta(\delta_\ell)$ la mesure de la distorsion relative entre δ_ℓ la sous-dominante ultramétrique δ_{n-1} de δ définie par : $\Delta(\delta_\ell) = \dfrac{\|\delta_\ell - \delta_{n-1}\|}{\|\delta_\ell\|} = 1 - \dfrac{\|\delta_{n-1}\|}{\|\delta_\ell\|}$. Quand ℓ croit, $\Delta(\delta_\ell)$ décroit strictement. La courbe rouge de la Figure 1 ci-dessus représente $\Delta(\delta_\ell)$ en fonction de ℓ lorsque δ est la distance euclidienne calculée sur les données Iris. La décroissance rapide $\Delta(\delta_\ell)$ lorsque ℓ croit, i.e. pour $\ell = 5$ la distorsion relative est environ 3%, suggère que δ_ℓ est presque ultramétrique, et par conséquent que les classes des hiérarchies d'Apresjan ne sont pas nécessairement de petites tailles, contrairement au cas général où δ est une dissimilarité quelconque (cf. Bryant et Berry (2001)). Néanmoins, pour évaluer la distorsion entre le dendrogramme, noté $\mathcal{D}_{\mathrm{SL}}(\delta)$, représentant la hiérarchie du lien simple $\mathcal{H}_{\mathrm{SL}}(\delta)$ et le dendrogramme $\mathcal{D}_\ell(\delta)$ représentant la hiérarchie d'Apresjan de δ_ℓ (indicée par la restriction de la fonction indice de la hiérarchie du lien simple), il est plus crédible d'utiliser la distorsion $\Delta(\rho_\ell)$ où ρ_ℓ désigne l'ultramétrique cophénétique induite par le dendrogramme $\mathcal{D}_\ell(\delta)$. Dans le cas des données Iris, la Figure 1 montre que la décroissance de $\Delta(\rho_\ell)$ (courbe bleue) est moins rapide que celle de $\mathcal{D}_\ell(\delta)$. Son graphe montre que les classes dont la longueur minimale de chemin est supérieure à 9 ne contribuent pas significativement à réduire la distorsion entre $\mathcal{D}_9(\delta)$ et la hiérarchie du lien simple de δ, puisque cette distorsion est déjà faible dès $\ell = 9$.

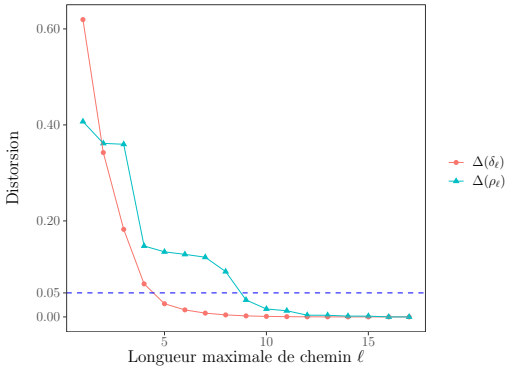


Figure 1: $\Delta(\delta_\ell)$ et $\Delta(\rho_\ell)$ en fonction de ℓ.

Bryant et Berry (2001) ont remarqué que les classes Apresjan sont généralement de petite taille. Cependant, la hiérarchie d'Apresjan d'une dissimilarité ultramétrique coïncide avec sa hiérarchie du lien simple, de sorte qu'elle peut contenir des classes de tailles arbitraires. Par un argument de continuité, la hiérarchie d'Apresjan d'une dissimilarité δ dont la distorsion $\Delta(\delta)$ est petite, peut également contenir des classes de taille arbitraire. D'après nos simulations, la distorsion $\Delta(\delta_\ell)$ diminue drastiquement dès que la valeur de ℓ augmente. Dans notre exemple, cela suggère que la hiérarchie d'Apresjan de δ_ℓ peut contenir des classes de taille arbitraire si $\ell \geq 3$. Si A est une classe d'Apresjan de δ_ℓ, alors pour tout $x \notin A$ et tout

$y, z \in A$, il existe un chemin P_{yz} de longueur au plus ℓ, tel que $\mathrm{val}_\ell(P_{yz}) < \delta(x, y)$. De plus, on peut montrer que si $val_\delta(P_{yz}) = \delta_\ell(y, z)$ alors le chemin P_{yz} est inclus dans A. Il est alors plausible de considérer comme valide chaque classe A de la hiérarchie du lien minimum dont la longueur de chemin minimale $\lambda(A)$ est à la fois petite et telle que $\lambda(A) \ll |A|$.

5 Quelques perspectives

Mentionnons brièvement quelques perspectives. D'après nos simulations, on peut observer que la suite $(\|\delta_\ell - \delta_{\ell+1}\|)_{1\leq\ell\leq n-1}$ est décroissante, lorsque ℓ croît. Cette propriété observée est à démontrer.
Par ailleurs, rappelons que $\ell_1 \leq \ell_2$ implique $\mathrm{conv}(J_{\ell_1}) \subseteq \mathrm{conv}(J_{\ell_2})$ (cf. théorème 3.2 (ii)). Plus généralement, comment définir les dissimilarités δ_ℓ à partir d'une suite de familles emboîtées de chemins valués par une dissimilarité δ, de sorte que cette propriété soit vraie ? Enfin, peut-on étendre notre approche à d'autres schémas de classification hiérarchique ?

Bibliographie

Apresjan, J. (1966) An algorithm for constructing clusters from a distance matrix, Mashinnyi perevod: prikladnaja lingvistika, 9, pp. 3-18.

Bandelt, H.-J., Dress, A.W.M. (1989) Weak hierarchies associated with similarity measures: an additive clustering technique, Bull. Math. Biol., 51, pp. 133-166.

Bertrand, P., Diatta, J. (2024) An interval convexity-based framework for multilevel clustering with applications to single-linkage clustering, Discrete Applied Mathematics, 342, pp. 38-63.

Bryant, D., Berry, V. (2001) A family of clustering and tree construction methods, Adv. Appl. Math., 27 (4), pp. 705-732.

Calder, J. (1971) Some elementary properties of interval convexities, J. Lond. Math. Soc. s2-3, 3, pp. 422-428.

Diatta, J., Fichet, B. (1998) Quasi-ultrametrics and their 2-Ball Hypergraphs, Discrete Math., 192, pp. 87-102.

Mirkin, B. and Muchnik, I. (1998) Combinatorial optimization in clustering, in: D.-Z. Du, P. Pardalos (Eds.), Handbook of Combinatorial Optimization, Kluwer Academic Publishers, Vol. 2, pp.261-329.

Van de Vel, M. (1993) Theory of Convex Structures, Elsevier, Amsterdam, North-Holland.

Xu, C., Liu, J., Tang, X. (2007) Clustering with transitive distance and K-means duality, arXiv:0711.3594.

Yu, Z., Liu, W., Liu, W., Yang, Y., Li, M., Vijaya Kumar, B.V.K. (2016) On order-constrained transitive distance clustering, in: AAAI, 2016.

Une approche divisive et interprétable de réduction de dimension pour la prédiction de la RUL

Mouhamadou Lamine NDAO [1,2] & Genane YOUNESS [1,2] & Ndèye NIANG [2] & Gilbert SAPORTA [2]

[1] *Laboratoire LINEACT CESI, Nanterre, IDFC*
[2] *Laboratoire Cédric-MSDMA-CNAM, Paris, France*
{mlndao ; gyouness}@cesi.fr ; {n-deye.niang ; gilbert.saporta}@cnam.fr

Résumé. En maintenance, la prédiction de la durée de vie utile restante (RUL) est entravée par la grande dimensionnalité et le manque d'explicabilité. Cette étude propose une approche combinant une méthode innovante de réduction de dimension interprétable, l'IDFC (Interpretable Divisive Feature Clustering), et un modèle LSTM (Long Short-Term Memory) à une couche. L'IDFC s'inspire des algorithmes VARCLUS (SAS) et CLV, identifiant efficacement les composantes latentes et les variables clés associées. Les expériences sur des données de roulements PRONOSTIA montrent que l'IDFC combiné à LSTM surpasse les méthodes classiques. La méthode XAI SHAP (SHapley Additive exPlanations) met en évidence les variables influençant la RUL, améliorant ainsi l'interprétabilité.

Mots-clés. Réduction de dimension, XAI, LSTM, classification de variables.

Abstract. Predicting Remaining Useful Life (RUL) involves challenges such as high dimension and lack of interpretability. This study introduces a new approach that combines Interpretable Divisive Feature Clustering (IDFC) with a single-layer LSTM (Long Short-Term Memory) model. IDFC, inspired by the VARCLUS and CLV algorithms, successfully identifies key variables. Experiments on PRONOSTIA bearing data show that IDFC combined with LSTM model outperforms traditional methods. Furthermore, SHAP(SHapley Additive exPlanations) method highlights the features influencing RUL, enhancing interpretability.

Keywords. Dimension reduction, XAI, LSTM, Feature clustering.

1 Introduction

La maintenance prédictive (MP) doit relever deux défis majeurs : la réduction de dimension face à la multitude de données industrielles et l'explicabilité des modèles complexes comme les LSTM pour la prédiction de la durée de vie restante (RUL). Les méthodes traditionnelles (Analayse en Composente Principale (ACP), Autoencodeurs, etc.) réduisent la dimension, mais génèrent des variables synthétiques difficiles à interpréter, limitant leur utilité pour une IA explicable (XAI). Des approches comme la classification de variables (Vigneau et al. 2003) améliorent l'interprétabilité, mais nécessitent souvent un nombre de classes prédéfini et peinent à clarifier les relations entre variables synthétiques et originales.

Cette étude propose une méthode de classification de variables interprétable basée sur un algorithme divisif (IDFC) visant à concilier performance prédictive et l'explicabilité. L'article est structuré comme suit : la Section 1 introduit le contexte et la problématique de l'étude, la Section 2 décrit la méthodologie ainsi que le protocole expérimental et la Section 3 présente les résultats, incluant une comparaison de la méthode proposée avec les approches traditionnelles.

2 Méthodologie et protocole expérimental

Nous désignons par $X = \{x^1, x^2, \ldots, x^p\}$ une matrice de p variables x^j décrivant N observations ; $\tilde{X} = \{c_1, c_2, \ldots, c_K\}$ une projection de X sur un espace latent de dimension réduite $K < p$ et $\mathcal{P}_\mathcal{K} = \{C_1, \ldots C_K\}$ une partition des variables en K classes.

2.1 Contexte

Il existe trois principales approches pour réduire la dimension de X : par projection, par sélection de variables, ou par une combinaison des deux. La troisième approche regroupe les variables en classes C_k, chacune représentée par une variable latente c_k résumant les variables de la classe associée (Vigneau et al. 2003 ; Chavent et al. 2012). Deux algorithmes sont couramment utilisés : VARCLUS (SAS) et CLV (Vigneau et al. 2003). VARCLUS divise les variables en deux classes en fonction de leur corrélation avec les deux premières composantes principales issues d'une ACP appliquée à l'ensemble des variables. Ce processus est ensuite itéré récursivement dans chaque classe jusqu'à ce que chaque classe C_k ait une seule valeur propre $\lambda_1^k > 1$, déterminant ainsi naturellement le nombre de classes K. La détermination du nombre de classes est ainsi inhérente à la méthode. La méthode CLV, de type kmeans, a pour objectif de trouver la partition $\mathcal{P}_K$ et les variables latentes $\tilde{X} = \{c_1, \ldots, c_K\}$ de manière à obtenir la valeur optimale T^* du critère CLV :

$$T^* = \max_{(\delta_{k,j}, c_k)} \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \operatorname{Cov}(x^j, c_k)^2, \tag{1}$$

sous la contrainte $\|c_k\| = 1$ avec $\delta_{kj} = 1$ si la variable x_j appartient au classe C_k, et $\delta_{kj} = 0$ sinon. Pour chaque classe C_k, c_k en est la première composante principale. Les deux algorithmes produisent des variables synthétiques $\tilde{X}$ non directement interprétables. Pour pallier ces problèmes, nous proposons une nouvelle approche de regroupement des variables, décrite dans la section suivante.

2.2 Interpretable Divisive Feature Clustering (IDFC)

IDFC intègre les avantages de VARCLUS et ceux de CLV. Il comprend trois étapes : une initialisation basée sur VARCLUS, un raffinement inspiré de CLV, et une extraction finale des variables interprétables.

Initialisation : Cette étape consiste à rechercher une partition initiale $\mathcal{P}_K$ à l'aide de VARCLUS. Cependant, VARCLUS peut potentiellement fournir un grand nombre de petites classes. Pour corriger cela, les classes avec une seule variable sont considérées comme des classes de bruit et sont fusionnées en une seule classe. Cette partition initiale est donnée par l'**Algorithme** 1.

Algorithm 1 Partition initiale

Entrée : $X = \{x^1, x^2, \ldots, x^p\}$.
Diviser les x^j de X en fonction des deux premières composantes obliques, notées C_1 et C_2 donnés par une ACP sur X.
for C_k **do**
    Diviser davantage C_k si $\lambda_2^k \geq 1$ en appliquant une ACP dans $X_{\mathbf{1}_{[\delta_{k,j}, j \in [1:p]]}}$.
    Pour chaque x^j, la réaffecter à C_k tel que $\max_{x^j \in C_k} r^2\left(x^j, c_k\right)$
    Arrêter la division lorsque $|C_k| = 1$ ou si $(\lambda_2^k < 1)$.
end for
Obtenir $\mathcal{P}_{\mathcal{K}}$ avec un nombre potentiellement élevé de classes contenant une seule variable.
Sortie : $\mathcal{P}_{\mathcal{K}'}$ avec $K' = K + 1$ classe fusionnant toutes les petites classes.

L'étape suivante consistera à utiliser une approche basée sur l'algorithme CLV pour consolider la partition initiale en donnant la possibilité de considérer les corrélations positives et négatives entre les variables.

Consolidation : Dans cette étape, la partition $\mathcal{P}_{K'}$ trouvée par l'**Algorithme** 1 est utilisée comme partition d'initialisation. Ensuite, un algorithme CLV est mis en place en considérant l'approche $K + 1$ proposée par Vigneau 2016. Il consiste à trouver T_{new} donné par :

$$T_{\text{new}} = \max_{\delta_{jk}, c_k} \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \operatorname{cov}^2\left(x^j, c_k\right) + \sum_{j=1}^{p} \left(1 - \sum_{k=1}^{K} \delta_{kj}\right) \rho^2 \operatorname{var}\left(x^j\right) \tag{2}$$

Pour chaque C_k, c_k en est la première composante principale. Dans l'équation 2, le paramètre ρ contrôle la classe de bruit choisi entre 0 et 1. Une valeur élevée de ρ conduit à affecter un nombre élevé de variables à la classe de bruit. Si $\rho = 0$, l'équation 1 sera considérée comme ne comportant pas de classe de bruit. Pour trouver la partition consolidée, l'algorithme suivant est proposé :

Algorithm 2 Raffinement avec CLV

Entrée : $X = \{x^1, x^2, \ldots, x^p\}$; $\mathcal{P}_{K'+1}$.
for Ck **do**
    Calculer la première composante principale c_k.
end for
for x^j **do**
    Affecter x^j à la classe maximisant le critère d'homogénéité interne T_{new}.
end for
Répéter jusqu'à convergence.
Sortie : la partition affinée $\mathcal{P}_{K'+1}$ et les composantes de chaque classe.

Extraction des variables interprétables Dans chaque classe, x^j la plus corrélée avec c_k est sélectionnée comme représentant de la classe :

$$x^k = \arg \max_{x^j \in C_k} r^2\left(x^j, c_k\right) \tag{3}$$

2.3 Expérimentations : prédiction de la RUL

L'objectif est d'identifier un modèle optimal f pour prédire la RUL. Le processus de modélisation consiste en acquisition des données, extraction de variables, prétraitement (normalisation, lissage), application d'une fenêtre temporelle (TW : Time Window) pour cibler la dégradation, réduction de dimension par IDFC, prédiction via un LSTM à une couche et interprétation des prédictions avec SHAP (Lundberg et al. 2017).

Les données utilisées proviennent des roulements de la plateforme PRONOSTIA (Nectoux et al. 2012), contenant des signaux d'accélération horizontaux et verticaux. À partir des vibrations, 111 variables sont extraites manuellement (Barraza et al. 2021). L'étude analyse trois conditions de fonctionnement, décrites dans le Tableau 1.

	Condition 1	**Condition 2**	**Condition 3**
Données d'entraînement	Bearing 1-1 Bearing 1-2	Bearing 2-1 Bearing 2-2	Bearing 3-1 Bearing 3-2
Données de test	Bearing 1-3 Bearing 1-4 Bearing 1-5 Bearing 1-6 Bearing 1-7	Bearing 2-3 Bearing 2-4 Bearing 2-5 Bearing 2-6 Bearing 2-7	Bearing 3-3

Table 1 – Description des données PRONOSTIA

3 Résultats et discussions

La Table 2 compare la méthode proposée (IDFC) à CLV et VARCLUS. Contrairement à CLV qui nécessite de fixer le nombre de classes à l'avance, VARCLUS et la IDFC le déterminent naturellement. Seuls CLV et IDFC distinguent les corrélations positives et négatives, tandis que VARCLUS et IDFC garantissent des classes unidimensionnels. En outre, IDFC génère des composantes parcimonieuses, n'est pas sensible à l'initialisation, et fournit des résultats interprétables, surpassant CLV et VARCLUS en flexibilité, gestion des corrélations et interprétabilité.

Critères	**CLV**	**VARCLUS**	**Méthode proposée : IDFC**
Nombre de classes K	fixé à l'avance	**non fixé à l'avance**	**non fixé à l'avance**
Unidimensionnalité	Non	**Oui**	**Oui**
Distingue les corrélations positives et négatives	**Oui**	Non	**Oui**
Composantes parcimonieuses	**Optionnel**	Non	**Optionnel**
Sensible à l'initialisation	Oui	**Non**	**Non**
Isolation des variables bruitées	**Oui**	Non	**Oui**
Résultat interprétable	Non	Non	**Oui**

Table 2 – Comparaison entre CLV, VARCLUS et la méthode proposée (IDFC)

Pour les Conditions 1, 2 et 3, 17, 22 et 20 variables sont identifiées comme bruit, respectivement. IDFC génère 33 classes pour les Conditions 1 et 2, et 31 pour la Condition 3, en plus d'une classe de bruit. La taille minimale d'une classe est de 2 (p. ex., C_3, C_{30}), tandis que certaines classes en contiennent davantage, comme C_3 (15 variables) et C_{24} (12 variables) pour la Condition 1.

Nous avons évalué l'efficacité de notre approche de réduction de dimension (IDFC combiné à LSTM) pour la prédiction de la RUL via une étude d'ablation, en la comparant à PCA combiné LSTM avec K composantes (K étant le nombre de variables sélectionnées par IDFC). La Table 3 montre que notre modèle surpasse légèrement PCA combiné à LSTM en précision. Par exemple, dans la Condition 1, notre modèle (IDFC + LSTM) obtient un RMSE de 0,084 et un MAE de 0,064, contre 0,098 et 0,074 pour PCA combiné à LSTM. Dans la Condition 2, le modèle proposé atteint un RMSE de 0,126 et un MAE de 0,103, contre 0,127 et 0,104 pour PCA combiné à LSTM, confirmant sa supériorité en se basant sur des variables originales.

Cond	Bearing	Méthode	RMSE	MAE
1	1-3	IDFC+LSTM	**0,063**	**0,051**
		PCA+LSTM	0,075	0,064
	1-4	IDFC+LSTM	0,071	0,056
		PCA+LSTM	**0,059**	**0,050**
	1-5	IDFC+LSTM	**0,151**	**0,121**
		PCA+LSTM	0,165	0,124
	1-6	IDFC+LSTM	**0,005**	**0,004**
		PCA+LSTM	0,012	0,010
	1-7	IDFC+LSTM	**0,133**	**0,088**
		PCA+LSTM	0,178	0,123
	Moyenne	IDFC+LSTM	**0,084**	**0,064**
		PCA+LSTM	0,098	0,074
		LSTM Qiu et al. 2006	0,152	0,120
		GRU Liu et al. 2023	0,181	0,147
		HDCN Zheng et al. 2022	0,232	0,198
		SOM+PCA+BiLSTM Liu et al. 2023	0,128	0,105
2	2-3	IDFC+LSTM	**0,039**	**0,030**
		PCA+LSTM	0,039	0,030
	2-4	IDFC+LSTM	0,037	0,031
		PCA+LSTM	**0,035**	**0,030**
	2-5	IDFC+LSTM	**0,243**	**0,195**
		PCA+LSTM	0,247	0,199
	2-6	IDFC+LSTM	**0,110**	**0,088**
		PCA+LSTM	0,113	0,090
	2-7	IDFC+LSTM	0,200	0,172
		PCA+LSTM	**0,198**	**0,169**
	Moyenne	IDFC+LSTM	**0,126**	**0,103**
		PCA+LSTM	0,127	0,104
		LSTM Qiu et al. 2006	0,270	0,219
		GRU Liu et al. 2023	0,226	0,171
		HDCN Zheng et al. 2022	0,283	0,244
		SOM+PCA+BiLSTM Liu et al. 2023	0,164	0,121
3	3-3	IDFC+LSTM	0,029	**0,001**
		PCA+LSTM	**0,028**	0,022

TABLE 3 – Performances par roulement et méthode, avec moyennes incluant d'autres approches de l'état de l'art

La Figure 3 montre l'efficacité du modèle proposé en comparant la RUL prédite et réelle du roulement 1_3. Les deux tendances coïncident étroitement, atteignant zéro en fin de période. Pour expliquer les résultats de notre modèle, nous avons utilisé SHAP pour identifier les variables influençant significativement la RUL prédite.

La méthode de réduction de dimension proposée permet d'appliquer des techniques de XAI pour expliquer les prédictions de la RUL prédite d'un modèle LSTM. Contrairement aux méthodes traditionnelles, elle produit des variables interprétables, donnant des résultats d'explication plus simples et directs. Avec SHAP, nous identifions les variables les plus influentes, comme energy_ban_6_freq_v et energy_ban_16_freq_v (Figure 3). Les méthodes classiques (PCA, t-SNE) génèrent des variables synthétiques non directement interprétables, limitant leur utilité pour l'explicabilité.

La Table 3 compare les performances de notre modèle à d'autres études utilisant le même jeu de données de PRONOSTIA. Notre approche surpasse légèrement les méthodes existantes,

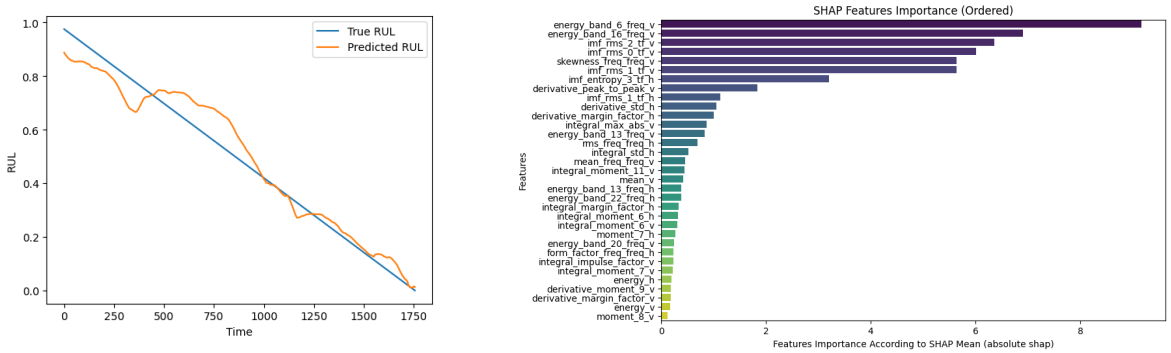


FIGURE 1 – Prediction de la RUL (à gauche) et l'importance des variables fournie par la méthode SHAP (à droite) : la somme des valeurs SHAP absolues pour les 4000 dernières secondes pour le bearing 1_3.

avec un RMSE moyen de 0,084 et un MAE moyen de 0,064 pour la condition 1, et un RMSE moyen de 0,126 et un MAE moyen de 0,103 pour la condition 3. De plus, notre modèle permet l'utilisation de méthodes XAI pour fournir des variables explicatives réelles, offrant une interprétabilité supérieure.

Références

BARRAZA, Joaquín Figueroa et al. (2021). « Towards Interpretable Deep Learning: A Feature Selection Framework for Prognostics and Health Management ». In : *Case Studies on Health Diagnostics and RUL Prediction* 12, p. 45-67.

CHAVENT, Marie et al. (2012). « ClustOfVar: An R package for the clustering of variables ». In : *Journal of Statistical Software* 50, p. 1-16.

LIU, Jie et al. (2023). « A Feature Fusion-Based Method for Remaining Useful Life Prediction of Rolling Bearings ». In : *IEEE Transactions on Instrumentation and Measurement*.

LUNDBERG, Scott M et al. (2017). « A unified approach to interpreting model predictions ». In : *Advances in neural information processing systems* 30.

NECTOUX, Patrick et al. (2012). « PRONOSTIA: An experimental platform for bearings accelerated degradation tests. » In : *IEEE International Conference on Prognostics and Health Management, PHM'12*. IEEE Catalog Number: CPF12PHM-CDR, p. 1-8.

QIU, Hai et al. (2006). « Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics ». In : *Journal of sound and vibration* 289.4-5, p. 1066-1090.

VIGNEAU, Evelyne (2016). « Dimensionality reduction by clustering of variables while setting aside atypical variables ». In : *Electronic Journal of Applied Statistical Analysis* 9.1, p. 134-153.

VIGNEAU, Evelyne et al. (2003). « Clustering of variables around latent components ». In : *Communications in Statistics-Simulation and Computation* 32.4, p. 1131-1150.

ZHENG, Likang et al. (2022). « Optimization of dilated convolution networks with application in remaining useful life prediction of induction motors ». In : *Measurement* 200, p. 111588.

Systèmes de Classes & Dissimilarités

Pascal Préa [1,2]

[1] *LIS, Aix-Marseille Université, CNRS, Marseille, France*
[2] *École Centrale Méditerranée, Marseille, France*
pascal.prea@lis-lab.fr

Résumé. Nous montrons ici les liens entre systèmes de classes et dissimilarités, en particulier pour les ultramétriques (liées aux hiérarchies), les dissimilarités de Robinson (liées aux pyramides) et les dissimilarités circulaires.

Mots-clés. Systèmes de classes, dissimilarités, ultramétriques, sériation, reconnaissance, approximation.

Abstract. We will show in this talk the links between class systems and dissimilarities, in particular for ultrametrics (and hierarchies), Robinson dissimilarities (and pyramids) and circular dissimilarities.

Keywords. Class systems, dissimilarities, ultrametrics, seriation, recognition, approximation.

1 Introduction

En reprenant G.L. Leclerc de Buffon (1749), on peut définir le but de la Classification comme *de mettre ensemble les choses qui se ressemblent et de séparer celles qui diffèrent les unes des autres*, ce qui implique d'une part de mettre les différentes entités dans différents ensembles (les *classes*) et d'autre part de mesurer la différence (ou la ressemblance) entre ces entités, ce que l'on fait souvent via une *dissimilarité* (une distance sans l'inégalité triangulaire – laquelle n'apporte rien ici). Le lien entre dissimilarités et systèmes de classes est donc un sujet fondamental en Classification, que nous allons expliciter.

La suite du document est organisée comme suit : Dans la section 2, nous définissons les systèmes de classes et donnons les principales manières de passer d'un système de classes à une dissimilarité et vice-versa. Dans la section 3, nous étudions les ultramétriques et les hiérarchies ; dans la section 4, les dissimilarités de Robinson et les pyramides, et enfin dans la section 5, les dissimilarités circulaires. Une conclusion conclura.

2 Définitions générales

Un *système de classes* sur un ensemble (fini) X est un sous-ensemble $\mathcal{C}$ de 2^X tel que :

$X, \emptyset, \{x\} \in \mathcal{C}$, $\forall x \in X$. Ces classes sont appelées *triviales*.

$\forall A, B \in \mathcal{C}$, $A \cap B \in \mathcal{C}$ ($\mathcal{C}$ est stable par intersection).

En rajoutant d'autres conditions, on obtient différents types de systèmes de classes. Le cas le plus simple est celui des partitions : un système de classes est une *partition* si deux classes non triviales ont toujours une intersection vide.

Une *dissimilarité* (*propre*) sur un ensemble X est une fonction $\mathbf{d} : X \times X \longmapsto \mathbb{R}$ telle que :

$\forall x, y \in X$, $\mathbf{d}(x, y) = \mathbf{d}(y, x) \geq 0$

$\mathbf{d}(x, y) = 0 \iff x = y$

À partir d'une dissimilarité, les deux algorithmes classiques pour obtenir une partition (le nombre de classes est fixé au départ) sont le *lien simple* (on maximise la somme des distances minimales entre classes) et le *lien moyen* (on maximise la somme des distances moyennes entre classes). Inversement, étant donné une partition $\mathcal{P}$, la dissimilarité, définie par (pour $x \neq y$) $\mathbf{d}(x, y) = 1$ si x et y sont dans une même classes non triviale et $\mathbf{d}(x, y) = 2$ sinon, est telle que $\mathcal{P}$ s'obtient à partir de $\mathbf{d}$ par le lien simple ou le lien moyen.

Nous allons nous intéresser ici à des systèmes de classes plus structurés que les partitions ; nous avons donc besoin de méthodes plus génériques pour passer d'une dissimilarité à un système de classes (et vice-versa).

Étant donné un système de classes $\mathcal{C}$, on définit la dissimilarité $\mathbf{d}_\mathcal{C}$ par $\mathbf{d}_\mathcal{C}(x, y)$ est (si $x \neq y$) le cardinal de la plus petite classe contenant x et y. Comme $X \in \mathcal{C}$ et $\mathcal{C}$ est stable par intersection, $\mathbf{d}_\mathcal{C}$ est bien définie.

Étant donné une dissimilarité $\mathbf{d}$, on peut définir des systèmes de classes en prenant l'ensemble des boules[1], des 2-boules[2] ou des cliques[3] de $\mathbf{d}$, puis en fermant cet ensemble par intersection.

3 Hiérarchies et Ultramétriques

Les hiérarchies et les ultramétriques sont les structures les plus classiques utilisées en classification et leur équivalence fait partie du folklore.

Un système de classes $\mathcal{C}$ est une *hiérarchie* si $\forall A, B \in \mathcal{C}$, $A \cap B \in \{A, B, \emptyset\}$, *i.e.* si, quand deux classes s'intersectent, une est incluse dans l'autre. La classification des êtres vivants est une hiérarchie.

Une dissimilarité est une *ultramétrique* si $\forall x, y, z \in X$, $\mathbf{d}(x, z) \leq \max\{\mathbf{d}(x, y), \mathbf{d}(y, z)\}$ (dans chaque triangle, les deux plus grands côtés sont de même longueur)

[1] Pour $x \in X$ et $\delta \geq 0$, la *boule* de centre x et de rayon δ est l'ensemble $B_\delta(x) := \{t \in X : \mathbf{d}(x, t) \leq \delta\}$.

[2] Pour $\{x, y\} \subset X$, la *2-boule* $B(x, y)$ est l'ensemble $B_{\mathbf{d}(x,y)}(x) \cap B_{\mathbf{d}(x,y)}(y) = \{t \in X : \mathbf{d}(x, y) \geq \max\{\mathbf{d}(x, t), \mathbf{d}(y, t)\}\}$.

[3] Une *clique* de rayon $\delta \geq 0$ est une clique maximale dans le graphe seuil (défini par δ) de $\mathbf{d}$, *i.e.* c'est un ensemble C tel que $\forall x, y \in C$, $\mathbf{d}(x, y) \leq \delta$ et $\forall z \in X \setminus C$, $\exists x \in C : \mathbf{d}(x, z) > \delta$.

Théorème 1 *Soit* $\mathbf{d}$ *une dissimilarité. Les assertions suivantes sont équivalentes :*

1. $\mathbf{d}$ *est une utramétrique.*
2. *Il existe une hiérarchie* $\mathcal{C}$ *telle que* $\mathbf{d} = \mathbf{d}_{\mathcal{C}}$.
3. *L'ensembles des boules de* $\mathbf{d}$ *est une hiérarchie.*
4. *L'ensembles des 2-boules de* $\mathbf{d}$ *est une hiérarchie.*
5. *L'ensembles des cliques de* $\mathbf{d}$ *est une hiérarchie.*

$\mathbf{d}$	a	b	c	d	e
a	0	1	4	4	4
b		0	4	4	4
c			0	3	3
d				0	2
e					0

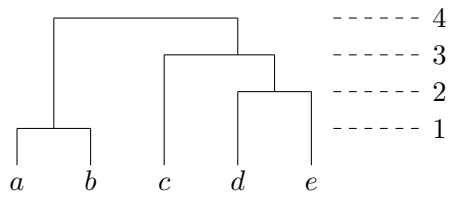


Figure 1: Une ultramétrique sur $X = \{a, b, c, d, e\}$ et sa hiérarchie associée. La hiérarchie est *indicée*, c'est à dire qu'à chaque classe est associée une hauteur. Pour $x, y \in X$, $\mathbf{d}(x, y)$ est la hauteur de la plus petite classe les contenant.

Reconnaitre les ultramétriques, c'est à dire déterminer si une dissimilarité sur un ensemble à n éléments est une ultramétrique ou pas peut se faire en temps optimal $O(n^2)$ (*cf.* Barthélemy & Brucker 2007). L'approximation (trouver l'ultramétrique la plus proche d'une dissimilarité donnée (ainsi que la hiérarchie associée)) est un problème NP-difficile en norme L_p avec p fini (Křivánek & Morávek 1986), mais polynomial en norme L_∞ (Chepoi & Fichet 2000) ; enfin, trouver l'ultramétrique sous-dominante (la plus grande ultramétrique plus petite qu'une dissimilarité donnée) est aussi polynomial (Gower & Ross 1969).

4 Pyramides, Sériation et Dissimilarités de Robinson

Généraliser les ultramétriques peut se faire en considérant des classes empiétantes. Le modèle standard pour de tels systèmes de classes est celui des pyramides. Un système de classes sur un ensemble X est une *pyramide* si X peut être ordonné de telle sorte que toutes les classes soient des intervalles (Diday 1986).

Une dissimilarité $\mathbf{d}$ sur un ensemble X est *Robinson* (1951) si il existe un ordre total (dit *compatible*) sur X tel que :

$$x < y < z \implies \mathbf{d}(x, z) \geq \max\{\mathbf{d}(x, y), \mathbf{d}(y, z)\}$$

Les dissimilarités de Robinson ont été inventées pour résoudre le problème de la sériation en archéologie (l'ordre compatible, que l'on cherche, est le temps) et sont maintenant un outil fondamental pour la sériation dans tous les domaines. Ce point de vue est plus orienté modélisation que classification. De plus, elles jouent un rôle important dans la détermination de cas polynomiaux pour le problème du voyageur de commerce (Çela & al. 2023, Deineko & al. 2014).

Proposition 1 (Diday 1986) *Les Ultramétriques sont des dissimilarités de Robinson. Les hiérarchies sont des pyramides.*

Théorème 2 (Durand & Fichet 1988) *Soit* $\mathbf{d}$ *une dissimilarité. Les assertions suivantes sont équivalentes :*

1. $\mathbf{d}$ *est Robinson.*
2. *Il existe une pyramide* $\mathcal{C}$ *telle que* $\mathbf{d} = \mathbf{d}_{\mathcal{C}}$.
3. *L'ensembles des boules de* $\mathbf{d}$ *clos par intersection est une pyramide.*
4. *L'ensembles des 2-boules de* $\mathbf{d}$ *clos par intersection est une pyramide.*
5. *L'ensembles des cliques de* $\mathbf{d}$ *clos par intersection est une pyramide.*

$\mathbf{d}$	a	b	c	d	e	f
a	0	1	3	3	5	5
b		0	3	3	5	5
c			0	1	2	3
d				0	2	3
e					0	1
f						0

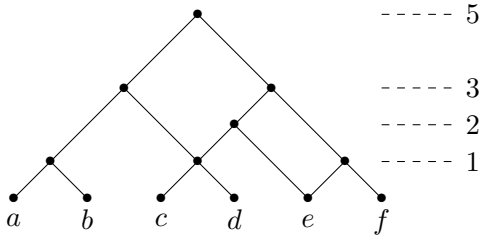


Figure 2: Une dissimilarité de Robinson ($a < b < c < d < e < f$ est un ordre compatible) et la pyramide associée. Comme sur la Figure 1, la pyramide est indicée et $\mathbf{d}(x, y)$ est la hauteur de la plus petite classe contenant x et y. On peut remarquer que, quand on s'éloigne de la diagonale, la matrice représentant $\mathbf{d}$ est non décroissante.

Il est possible de reconnaitre les dissimilarités de Robinson en temps optimal $O(n^2)$ (Préa & Fortin 2014, Carmona & al. 2024) mais l'approximation est NP-difficile, au moins pour les normes L_1 (Barthélemy & Brucker 2001) et L_∞ (Chepoi & al. 2009). Pour la norme L_∞, Chepoi & Seston (2011) ont donné un algorithme d'approximation de facteur 16. Par contre, l'approximation à ordre fixé (étant donné une dissimilarité $\mathbf{d}$ sur un ensemble X et un ordre total $<$ sur X, trouver la dissimilarité de Robinson la plus proche de $\mathbf{d}$ admettant $<$ comme ordre compatible) se résout en temps polynomial (ce problème se ramène aisément à la régression isotonique).

5 Dissimilarités Circulaires

Pour étendre les dissimilarités de Robinson, on peut considérer un cercle au lieu d'une droite. Un *ordre circulaire* est une relation sur trois points : $x < y < z$ signifie que x, y et z sont dans l'ordre trigonométrique direct. Cette écriture s'étend à $x_1 < x_2 < \ldots < x_p$. Les ordres circulaires ont été formalisés par Huntington en 1916. Si $<$ est un ordre circulaire, pour $x, y \in X$, l'*arc* $X^{<}_{xy}$ est l'ensemble $\{t : x < t < y\} \cup \{x, y\}$.

D'un point de vue "classification", on a trois définitions de circularité pour une dissimilarité $\mathbf{d}$: $\mathbf{d}$ est *quasi-circulaire* (resp. *2-boules-circulaires*, resp. *cliques-circulaire*) si toutes ses boules (resp. 2-boules, resp. cliques) sont des arcs.

D'un point de vue "modélisation", on dira qu'une dissimilarité $\mathbf{d}$ est *circulaire* avec un ordre compatible $<$ si, pour tous $x, y \in X$, un des deux arcs $X^{<}_{xy}$ et $X^{<}_{yx}$ est Robinson avec la restriction de $<$ comme ordre compatible.

Théorème 3 (Brucker & Osswald 2008, Carmona & al. 2023) *Une dissimilarité* $\mathbf{d}$ *est quasi-circulaire d'ordre compatible* $<$ *si et seulement si :*

$$\forall\, x < y < z < t,\ \mathbf{d}(x,z) \geq \min\{\mathbf{d}(y,z), \mathbf{d}(t,z)\}$$

Une dissimilarité $\mathbf{d}$ *est circulaire d'ordre compatible* $<$ *si et seulement si :*

$$\forall\, x < y < z < t,\ \mathbf{d}(x,z) \geq \min\{\max\{\mathbf{d}(x,y), \mathbf{d}(y,z)\}, \max\{\mathbf{d}(x,t), \mathbf{d}(t,z)\}\}$$

Toutes ces dissimilarités généralisent les dissimilarités de Robinson, mais contrairement au cas linéaire, on a :

Théorème 4 (Brucker & Osswald 2008, Carmona & al. 2023) *Soit* $\mathbf{d}$ *une dissimilarité. On a* (1) $\Longrightarrow$ (2) $\Longrightarrow$ (3) $\Longrightarrow$ (4) $\Longrightarrow$ (5)*, où :*

1. $\mathbf{d}$ *est Robinson.*
2. $\mathbf{d}$ *est cliques-circulaire.*
3. $\mathbf{d}$ *est 2-boules-circulaire.*
4. $\mathbf{d}$ *est circulaire.*
5. $\mathbf{d}$ *est quasi-circulaire.*

Les réciproques sont fausses.

	Cas Général	Cas Strict
Quasi-circulaires	$O(n^3)$ [2]	$O(n^2)$ [1, 3]
Circulaires		$O(n^2)$ [3]
2-boules-Circulaires	$O(n^3)$ [2]	$\downarrow$
Cliques-Circulaires	$O(n^3)$ [2]	$\rightarrow$

Table 1: Complexités des meilleurs algorithmes de reconnaissance pour les différentes définitions de la circularité. Le cas strict se définit en remplaçant les "$\geq$" du théorème 3 par des "$>$" (il existe des formules similaires pour les dissimilarités 2-boules ou cliques-circulaires). La référence [1] correspond à Armstrong & al. 2021, [2] à Brucker & Osswald 2008 et [3] à Carmona & al. 2023. Les algorithmes pour le cas général s'appliquent au cas strict et l'algorithme pour reconnaitre les dissimilarités circulaires strictes peut s'adapter aux 2-boules-circulaires strictes.

6 Conclusion

L'étude des dissimilartités établit un lien entre Classification et Modélisation et les dissimilarités de Robinson occupent de ce point de vue une place privilégiée. Elles se généralisent à toutes les définitions de "betweeness", en disant que quand y est "entre" x et z, alors $\mathbf{d}(x,z) \geq \max\{\mathbf{d}(x,y), \mathbf{d}(y,z)\}$. De très nombreux problèmes restent ouverts, par exemple, trouver un algorithme polynomial pour reconnaitre les dissimilarités circulaires dans le cas général.

Bibliographie

Armstrong, A, Guzmán, C. & Sing Long, C.A. (2021), An optimal algorithm for strict circular seriation, *SIAM Journal on Mathematics of Data Science* 3(4), pp.1223-1250.

Barthélemy, J-P. & Brucker, B. (2001), NP-hard approximation problems in overlapping clustering, *J. Classif.* 18, pp. 159-183 .

Barthélemy, J-P. & Brucker, F. (2007), *Éléments de Classification*, Hermès.

Brucker B. & Osswald, C. (2008), Hypercycles and dissimilarities, *J. Classif*, accepté.

Carmona, M, Chepoi, V, Naves, G. & Préa, P. (2023), A Simple and Optimal Algorithm for Strict Circular Seriation, *SIAM Journal on Mathematics of Data Science* 5(1), pp. 201-221.

Carmona, M, Chepoi, V, Naves, G. & Préa, P. (2024), Modules in Robinson Spaces, *SIAM Journal on Discrete Mathematics* 38, pp.190-224.

Çela, E, Deineko, V. & Woeginger, G.J. (2023), Recognising permuted Demidenko matrices, *Operations Research Letters*, 51, pp 494-500.

Chepoi, V. & Fichet, B. (2000), L_∞-approximation via subdominants, *Journal of mathematical psychology*, 44, pp. 600-616.

Chepoi, V, Fichet, B. & Seston, M. (2009), Seriation in the presence of errors: NP-hardness of l_∞-fitting Robinson structures to dissimilarity matrices, *J. Classif.* 26, pp. 279-296.

Chepoi, V. & Seston, M. (2011), Seriation in the presence of errors : A factor 16 approximation algorithm for l_∞-fitting Robinson structures to distances, *Algorithmica* 59, pp. 521-568.

Deineko, V, Klinz, B, Tiskin, A. & Woeginger, G.J. (2014), Four-point conditions for the TSP : The complete complexity classification, *Discrete optimization* 14, pp. 147-159.

Diday, E. (1986), Orders and overlapping clusters in Pyramids, in *Multidimensional Data Analysis* De Leew & al. Eds, pp. 201-234.

Durand, C. & Fichet, B. (1988), One-to-One Correspondences in Pyramidal Representation: An Unified Approach, in *Classification and Related Methods of Data Analysis*, H.H. Bock Ed., pp. 85–90.

Gower, J.C. & Ross, G.J.S. (1969), Minimal spanning trees and single linkage cluster analysis, *Applied Statistics*, 18, pp. 54-64.

Huntington, E.V. (1916), A set of independent postulates for cyclic order, *Proceedings of the National Academy of Sciences of the United States of America* 2(11), p. 630.

Křivánek, M. & Morávek, J. (1986), NP-hard problems in hierarchical-tree clustering, *Acta Informatica*, 23, pp. 311-323.

Leclerc de Buffon, G.L. (1749), *Histoire Naturelle.*

Préa, P. & Fortin, D. (2014), An optimal algorithm to recognize Robinsonian dissimilarities, *J. Classif.* 31, pp. 351-385.

Robinson, W.S. (1951), A method for chronologically ordering archaeological deposits, *Amer. Antiq.* 16, pp. 293-301.

Stat & Sport (session du groupe)

Clustering of soccer possessions using mixture of Hawkes processes with Markov marks

Koffi AMEZOUWUI [1] & Brigitte GELEIN [2] & Matthieu MARBAC [1] & Anthony SOREL [3]

[1] *Univ. Rennes, Ensai, CNRS, CREST-UMR 9194, F-35000 Rennes, France, koffi.amezouwui@ensai.fr, matthieu.marbac-lourdelle@ensai.fr*
[2] *Ensai, IRMAR-UMR, CNRS 6625, France, brigitte.gelein@ensai.fr*
[3] *Univ. Rennes, Inria, M2S, F-35000 Rennes, France, anthony.sorel@univ-rennes2.fr*

Résumé:

On s'intéresse à la classification non supervisée de phases de possession de football décrites par un processus ponctuel marqué, caractérisant les temps d'arrivée des actions et les types d'actions. Pour cela, nous développons un modèle de mélange de processus de Hawkes marqués, dont les marques, modélisant les types d'actions, suivent une chaîne de Markov. L'estimation des paramètres se fait par maximum de vraisemblance au moyen d'un algorithme Generalized Expectation-Maximization (GEM). Nos résultats offrent des perspectives prometteuses pour l'analyse de performance et l'aide à la décision en football.

Mots-clés: Processus de Hawkes, Chaînes de Markov, Clustering, Modèles de mélange, Football.

Abstract: We focus on the unsupervised classification of football possession phases, described by a marked point process that characterizes the arrival times of actions and their types. To this end, we develop a mixture model of marked Hawkes processes, where the marks (modeling the action types) follow a Markov chain. Parameter estimation is performed via maximum likelihood using a Generalized Expectation-Maximization (GEM) algorithm. Our results offer promising perspectives for performance analysis and decision support in football.

Keywords: Hawkes processes, Markov chains, Clustering, Mixture models, Football.

1 Introduction

In this paper, we address the problem of clustering soccer game possessions using events recorded by tracking systems during a game. The analysis of possession sequences in soccer has become a critical tool for understanding game dynamics, revealing intricate patterns of play, and informing tactical decision-making, with implications for player development and competitive success (Hughes et al., 2019). However, the complex, dynamic, and multi-scale nature of soccer, characterized by nonlinear interactions between players and the ball, presents substantial challenges for traditional analytical approaches, which often struggle to capture contextual information, underlying dependencies, and emergent behavior inherent

in team sports (Link et al., 2016; Goes et al., 2019). The data to be clustered consist of football possession phases, described by a marked point process that characterizes the arrival times of actions and their types. Clustering can be achieved by modeling the distribution of the observed data with a mixture model. In such a case, clusters can be defined as subsets of the possession phases arising from the same mixture component. In this context, the distribution of each component depends on the nature of the data. Hence, to cluster the point process describing the movement patterns of bats, Denis et al., (2023) developed a mixture of Hawkes processes. This model is relevant for fitting the processes describing the arrival times of actions, since the self-exciting property of the Hawkes process allows for rhythm variation within the same possession phases. However, this model cannot directly take into account the marked events.

To address this issue, we propose an approach based on a mixture of marked Hawkes processes with discrete Markov chains as marks. This methodological framework is particularly well-suited for football analysis, as it enables the modeling of both temporal dependencies and strategic transitions within phases of possession, thereby capturing the influence of past actions on future events. By clustering these sequences into meaningful groups, it becomes possible to characterize different playing styles, predict the evolution of possession phases, and refine team strategies. The combination of marked Hawkes processes, which represent the temporal dependencies between events, and Markov chains, which describe sequential transitions, provides a structured and efficient framework for analyzing ball possession. However, this approach requires precise parameter estimation, appropriate inference techniques, and rigorous validation to ensure its robustness and interpretability in real-game scenarios.

This paper is organized as follows. Section 2 introduces the new mixture of Hawkes processes with Markov marks. Section 3 presents the GEM algorithm used for maximum likelihood estimation. Section 4 presents numerical experiments on simulated data. Section 5 concludes with insights and future research directions.

2 Mixture model of Hawkes process with Markov marks

Let $\mathbf{S} = \{\boldsymbol{S}_i\}_{i=1}^n$ denote a collection of n independent and identically distributed sequences of events, where $\boldsymbol{S}_i = (\boldsymbol{S}_{i1}, \ldots, \boldsymbol{S}_{in_i})$ describes the n_i events of sequence i. Event j of possession i is described by $T_{ij} \in \mathbb{R}^+$, the time of arrival of event j since the possession i started, and the type of event $U_{ij} \in \mathcal{E}$ where $\mathcal{E} = \{1, \ldots, E\}$, leading to $\boldsymbol{S}_{ij} = (T_{ij}, U_{ij})$, where E is the number of events. To model heterogeneity in the sequences, for each sequence $\boldsymbol{S}_i$, the variable $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iK})$ represents the cluster assignment for possession i, where $Z_{ik} = 1$ if possession i belongs to cluster k, and $Z_{ik} = 0$ otherwise, K being the number of clusters. The latent variables $\boldsymbol{Z}_i$ are independent and follow a multinomial distribution parameterized by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, where $\pi_k = \mathbb{P}(Z_{ik} = 1)$ denotes the probability of a sequence belonging to cluster k, satisfies $\sum_{k=1}^K \pi_k = 1$. The model considers that $\boldsymbol{T}_i = (T_{i1}, \ldots, T_{in_i})$ and $\boldsymbol{U}_i = (U_{i1}, \ldots, U_{in_i})$ are conditionally independent given $\boldsymbol{Z}_i$. We suppose that the conditional distribution of $\boldsymbol{T}_i$ given $\boldsymbol{Z}_i$ follows a Hawkes process and $\boldsymbol{U}_i$ (the type of events i.e., the marks) given $\boldsymbol{Z}_i$ follows a finite Markov chain, whose parameters are defined according to $\boldsymbol{Z}_i$. Therefore, the marginal density of $\boldsymbol{S}_i$ is defined by

$$f(\boldsymbol{s}_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k g(\boldsymbol{u}_i;\boldsymbol{\Gamma}_k,\boldsymbol{\nu}_k) h(\boldsymbol{t}_i;\boldsymbol{\rho}_k), \tag{1}$$

with $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_K, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K, \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_K)$. The function $h(\cdot;\boldsymbol{\rho}_k)$ denotes the density of Hawkes process with parameter $\boldsymbol{\rho}_k$ and $g(\cdot;\boldsymbol{\Gamma}_k,\boldsymbol{\nu}_k)$ denotes the density probability function of a Markov chain parametrized by the transition matrix $\boldsymbol{\Gamma}_k$ and the initial distribution $\boldsymbol{\nu}_k$. Hence, we have

$$g(\boldsymbol{u}_i;\boldsymbol{\Gamma}_k,\boldsymbol{\nu}_k) = \boldsymbol{\nu}_k(u_{i1}) \prod_{j=2}^{n_i} \boldsymbol{\Gamma}_k[u_{ij-1}, u_{ij}].$$

We consider a Hawkes process with an exponential decay excitation function such that

$$h(\boldsymbol{t}_i;\boldsymbol{\rho}_k) = \left[\prod_{j=1}^{n_i} \lambda\left(t_{ij} \mid \mathcal{H}_i(t_{ij});\boldsymbol{\rho}_k\right)\right] \exp\left(-\Lambda(t_{in_i};\boldsymbol{\rho}_k)\right),$$

where the intensity function is defined by $\lambda(t \mid \mathcal{H}_i(t);\boldsymbol{\rho}_k) = \mu_k + \sum_{t_{ij}<t} \alpha_k e^{-\beta_k(t-t_{ij})}$ with $\boldsymbol{\rho}_k = (\mu_k, \alpha_k, \beta_k) \in \mathbb{R}_+^{*3}$. The compensator is $\Lambda\left(t_{in_i};\boldsymbol{\rho}_k\right) = \mu_k t_{in_i} - \frac{\alpha_k}{\beta_k} \sum_{j=1}^{n_i} \left(e^{-\beta_k(t_{in_i}-t_{ij})} - 1\right)$.

3 Maximum likelihood inference

We proposed to estimate the model parameters $\boldsymbol{\theta}$ by maximizing the log-likelihood function of the model defined by $\ell(\boldsymbol{\theta};\mathbf{S}) = \sum_{i=1}^{n} \ln\left(\sum_{k=1}^{K} \pi_k g(\boldsymbol{u}_i;\boldsymbol{\Gamma}_k,\boldsymbol{\nu}_k) h\left(\boldsymbol{t}_i;\boldsymbol{\rho}_k\right)\right)$. This maximization is achieved via an EM algorithm Dempster (1977) which considers the completed-data log-likelihood defined by

$$\ell_c(\boldsymbol{\theta};\mathbf{S},\mathbf{Z}) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left[\ln \pi_k + \ln g(\boldsymbol{u}_i;\boldsymbol{\Gamma}_k,\boldsymbol{\nu}_k) + \ln h\left(\boldsymbol{t}_i;\boldsymbol{\rho}_k\right)\right].$$

The EM algorithm is an iterative algorithm randomly initialized at the model parameter $\boldsymbol{\theta}^{[0]}$. It alternates between two steps: the expectation step (E-step) which consists of computing $Q(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}})$ defined by $Q(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{Z}|\mathbf{S},\tilde{\boldsymbol{\theta}}}\left[\ell_c(\boldsymbol{\theta};\mathbf{S},\mathbf{Z})\right]$, the expectation of the complete-data likelihood under the current parameters $\tilde{\boldsymbol{\theta}}$. The maximization step (M-step), which consists of maximizing this expectation with respect to the model parameters. However, it is not always necessary to fully maximize $Q(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}})$; instead, it is sufficient to ensure that it increases at each M-step. For this reason, we adopt the Generalized Expectation-Maximization (GEM) algorithm, which reduces the computational complexity of exact maximization while still ensuring convergence. Iteration $[r]$ of the algorithm is defined by:

- E-step: we compute $r_{ik}(\boldsymbol{\theta}^{[r]})$ the conditional probability that sequence i belongs to cluster k, given the observed data and current parameter estimates

$$r_{ik}(\boldsymbol{\theta}^{[r]}) = \frac{\pi_k^{[r]} h\left(\mathbf{t}_i;\boldsymbol{\rho}_k^{[r]}\right) g\left(\boldsymbol{u}_i;\boldsymbol{\Gamma}_k^{[r]},\boldsymbol{\nu}_k^{[r]}\right)}{\sum_{k'=1}^{K} \pi_{k'}^{[r]} h\left(\mathbf{t}_i;\boldsymbol{\rho}_{k'}^{[r]}\right) g\left(\boldsymbol{u}_i;\boldsymbol{\Gamma}_{k'}^{[r]},\boldsymbol{\nu}_{k'}^{[r]}\right)}.$$

- M-step : parameter updating of $\pi_k^{[r+1]} = \dfrac{\sum_{i=1}^{n} r_{ik}(\boldsymbol{\theta}^{[r]})}{n}$; $\boldsymbol{\nu}_k^{[r+1]}(u) = \dfrac{\sum_{i=1}^{n} r_{ik}(\boldsymbol{\theta}^{[r]})\mathbb{1}_{(u_{i1}=u)}}{n}$, $\boldsymbol{\Gamma}_k[u,v]^{[r+1]} = \dfrac{\sum_{i=1}^{n}\sum_{j=2}^{n_i} r_{ik}(\boldsymbol{\theta}^{[r]})\mathbb{1}_{(u_{ij-1}=u,u_{ij}=v)}}{\sum_{v'\in\mathcal{E}}\sum_{i=1}^{n}\sum_{j=2}^{n_i} r_{ik}(\boldsymbol{\theta}^{[r]})\mathbb{1}_{(u_{ij-1}=u,u_{ij}=v')}}$; $\boldsymbol{\rho}_k^{[r+1]} = \boldsymbol{\rho}_k^{[r]} - \mathbf{H}_k(\boldsymbol{\theta}^{[r]})^{-1}\nabla\mathbf{G}_k(\boldsymbol{\theta}^{[r]})$ for each pair of states $(u,v) \in \mathcal{E}^2$ and cluster k.

The gradients $\nabla\mathbf{G}_k(\boldsymbol{\theta})$ and the Hessian matrix $\mathbf{H}_k(\boldsymbol{\theta})$ respect to $\boldsymbol{\rho}_k$ required for the Newton-Raphson updates method are obtained following Ozaki (1979).

4 Numerical experiments

4.1 Simulation design

All simulations are conducted using ($K = 2$), using a mixture of marked Hawkes processes and finite Markov chains, with $\boldsymbol{\pi} = (0.3, 0.7)$ representing the class membership probabilities. The Hawkes process is characterized by its parameters $\boldsymbol{\rho}_k = (\mu_k, \alpha_k, \beta_k)$, while the dynamics of the Markov chains are defined by the initial distributions $\boldsymbol{\nu}_k$ and transition matrices $\boldsymbol{\Gamma}_k$ for each k. We generate a set of n possession sequences, denoted as $\boldsymbol{S} = \{\boldsymbol{S}_i\}_{i=1}^{n}$ where each sequence $\boldsymbol{S}_i = (\boldsymbol{T}_i, \boldsymbol{U}_i)$. $\boldsymbol{T}_i$ follows an exponential decay Hawkes process with parameters $\boldsymbol{\rho}_1 = (1.5, 0.5, 1.0)$ if $Z_{i1} = 1$, and $\boldsymbol{\rho}_2 = (3.0, 0.5, 1.0)$ if $Z_{i2} = 1$. $\boldsymbol{U}_i$ folows a finite Markov chain with $E = 3$ states, characterized by their initial distributions $\boldsymbol{\nu}_1 = \boldsymbol{\nu}_2 = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$ and transition matrices $\boldsymbol{\Gamma}_1 = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$ if $Z_{i1} = 1$, $\boldsymbol{\Gamma}_2 = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$ if $Z_{i2} = 1$.

Let $\mathcal{I} = \left\{(\boldsymbol{\pi}^{[l]}, \boldsymbol{\rho}_k^{[l]}, \boldsymbol{\nu}_k^{[l]}, \boldsymbol{\Gamma}_k^{[l]}); l = 1, \ldots, m\right\}$ with $\boldsymbol{\rho}_k^{[l]} = \left\{(\mu_k^{[l]}, \alpha_k^{[l]}, \beta_k^{[l]}); k = 1, \ldots, K\right\}$ be the set of $m = 10$ initializations, such that $\boldsymbol{\pi}^{[l]}$follow a Dirichlet distribution $\mathcal{D}(\mathbf{1}_K)$ and $\mathbf{1}_K$ (a vector of size K filled with ones).

$\forall k \in \{1, \ldots, K\}$, $\boldsymbol{\Gamma}_k^{[l]} = \frac{1}{E}\mathbb{1}_{E\times E}$ is a matrix of size $E \times E$ with uniform values $\frac{1}{E}$ and $\boldsymbol{\nu}_k^{[l]} = \frac{1}{E}\mathbb{1}_{K\times E}$ the initial distribution; $\mu_k = \frac{s_T}{T}$ where s_T is drawn randomly over the number of events n_i then α_k and β_k follow a uniform distribution $\mathcal{U}(0,1)$,$\mathcal{U}(0,5)$ respectively.

4.2 Results

The Adjusted Rand Index (ARI) is a clustering evaluation metric that measures the similarity between two data clusterings while adjusting for chance. This metric is widely used in clustering analysis to assess the performance of unsupervised learning algorithms. It is based on the Rand Index (RI), which calculates the proportion of correctly grouped elements (both in the same and different clusters) in two partitions.

To illustrate the convergence of the estimators when the model parameters are estimated using maximum likelihood, we compute the adjusted Rand index (ARI) (Hubert and Arabie, 1985) between the true partition and the partition obtained from the maximum a posteriori probability. Table 1 presents the results for different sample sizes ($n = 50, 100, 200, 400$) over $R = 50$ replications. This illustrates the convergence of the estimators when replicates are drawn, measuring the ARI between the estimated and true partitions.

	ARI (sd)			
Sample sizes (n) / Methods	50	100	200	400
Mixture of HP & MC	**0.74** (0.25)	**0.80** (0.17)	**0.81** (0.15)	**0.82** (0.11)
Mixture of HP	0.33 (0.36)	0.30 (0.33)	0.39 (0.35)	0.28 (0.33)
Mixture of MC	0.51 (0.20)	0.60 (0.14)	0.59 (0.12)	0.60 (0.08)

Table 1: Mean adjusted Rand Index (ARI) computed over $R = 50$ replications for different sample sizes n. Each ARI is reported with its standard deviation (sd). Note that (HP) stands for Hawkes processes and (MC) for Markov chain.

The results of Table 1 show that the method combining Hawkes Processes and Markov chains consistently achieves the highest Adjusted Rand Index (ARI), which increases as the sample size grows, indicating an improvement in the alignment between the estimated and true partition. This suggests that the estimator benefits from larger datasets, leading to more accurate partitioning. In contrast, The Hawkes processes approach shows lower and less stable performance, while the Markov chains method achieves intermediate results with ARI improving as the sample size increases. The ARI values also depend on the separability of the classes. For example, the more different the $\boldsymbol{\rho}_k$ are from each other, the easier it is for the algorithm to distinguish them. More extensive simulations are planned.

5 Conclusion

In the present work, we introduced a novel clustering approach for soccer game possessions, leveraging a mixture of Hawkes processes marks with Markov chains. This methodological framework effectively captures both temporal dependencies and strategic transitions, addressing the inherent limitations of traditional clustering techniques that struggle to incorporate the complexity of the game. Our results indicate that the Adjusted Rand Index (ARI) increases with sample size, demonstrating that the proposed approach benefits from larger datasets, leading to more accurate partitioning. This underscores the potential of our method in identifying recurrent patterns of play and anticipating the evolution of possession sequences. Nevertheless, despite these improvements, several challenges remain, particularly concerning parameter estimation, model interpretability, and computational efficiency. Future research should focus on refining the inference process, integrating additional contextual variables, spatial aspects and extending the model to real-time applications for tactical decision-making. Ultimately, this study highlights the crucial role of advanced statistical models in sports analytics, laying the foundation for more sophisticated approaches that

enhance performance evaluation, optimize training strategies, and strengthen data-driven decision-making in soccer and beyond.

Bibliographie

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society*: series B (methodological), 39(1), 1-22.

Denis, C., Dion-Blanc, C., Lacoste, R. E., Sansonnet, L., Bas, Y. (2023). Bats monitoring: a classification procedure of bats behaviours based on Hawkes processes. *Journal of the Royal Statistical Society Series C: Applied Statistics*.

Goes, F. R., Bastos, F. C., Aquino, R., Folgado, H., and Travassos, B. (2019). Identifying Tactical Patterns of Play in Elite Soccer Using Positional Data. *Frontiers in Psychology,*10.

Hubert, L., Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2, 193-218.

Hughes, M., Franks, I., Hughes, M., Franks, I.M., Dancs, H. (Eds.). (2019). Essentials of Performance Analysis in Sport: Third edition (3rd ed.). Routledge.

Link, D., Langguth, M., and Wahl, F. M. (2016). *Real-time motion capture and analysis of players' trajectories in sports. Computer Graphics Forum*, 35(2), 527–538.

Ozaki, T. (1979). Maximum likelihood estimation of hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1) :145– 155.

Application des Modèles de Markov dérivants à la dynamique d'acquisition des compétences

Ludovic Seifert[1,4] & Emmanouil-Nektarios Kalligeris [1,2,5] & John Komar [6] & Vlad Stefan Barbu [2,3] & Guillaume Hacques [1] & Nicolas Vergne [2]

[1] *Centre d'études des Transformations des Activités Physiques et Sportives, Université de Rouen Normandie, Boulevard Siegfried, 76821 Mont-Saint-Aignan, Rouen, France, guillaume.hacques@univ-rouen.fr*

[2] *Laboratoire de Mathématique Raphaël Salem, Université de Rouen Normandie, Avenue de l'Université, BP.12, 76801 Saint-Étienne du Rouvray, Rouen, France, nicolas.vergne@univ-rouen.fr*

[3] *Centre for Demographic Research "Vladimir Trebici", "Costin C. Kiritescu" National Institute of Economic Research of Romanian Academy, Romania, barbu@univ-rouen.fr*

[4] *Institut Universitaire de France (IUF), Paris, France, ludovic.seifert@univ-rouen.fr*

[5] *Circana, Greece, Manolis.Kalligeris@Circana.com*

[6] *Nanyang Technological University, Physical Education and Sport Science, National Institue of Education, Singapore, john.komar@nie.edu.sg*

Résumé. Ce travail explore le cadre multidisciplinaire de la dynamique écologique appliquée au comportement humain et présente les modèles de Markov dérivants. Tout d'abord, les connaissances les plus récentes et les défis liés à la dynamique de l'acquisition des compétences sont brièvement résumés. Ensuite, la contribution innovante du cadre de la dynamique écologique à la dynamique de l'acquisition des compétences et la valeur ajoutée des modèles de Markov dérivants sont exposées. Enfin, à partir de deux études, nous montrons comment les modèles de Markov dérivants peuvent mieux expliquer et comprendre la dynamique de l'acquisition des compétences, en particulier l'alternance entre l'*exploration* de nouvelles solutions motrices et l'*exploitation* de solutions motrices existantes ou récemment stabilisées dans le répertoire de l'apprenant, un phénomène appelé le régime métastable.

Mots-clés. Modèles de Markov dérivants, Dynamique persistante, Coordination visuo-motrice

Abstract. This work explores the multidisciplinary framework of Ecological Dynamics as applied to human behavior and outlines the drifting Markov models. First, the most up-to-date knowledge and challenges related to the dynamics of skill acquisition are briefly summarized. Next, the innovative contribution of the Ecological Dynamics framework to skill acquisition dynamics and the added value of drifting Markov models are presented. Finally, based on two studies, we demonstrate how drifting Markov models can better explain and understand the dynamics of skill acquisition, particularly the alternation between "exploration" of new motor solutions and "exploitation" of existing or recently stabilized motor solutions in the learner's repertoire, a phenomenon referred to as the metastable regime.

Keywords. Drifting Markov Models, Persistent Dynamics, Visual-Motor Coordination

1 La dynamique de l'acquisition des compétences

Le cadre de la dynamique écologique de l'acquisition des compétences suggère que ce processus implique des transitions entre états stables, menant soit à l'émergence de nouveaux motifs comportementaux, soit à la stabilisation de motifs existants (Button et al., 2020). L'acquisition d'une compétence repose sur la stabilité des comportements, définie comme la capacité à revenir rapidement à un état initial après perturbation (Haken, 1983). La dégénérescence dans les systèmes neurobiologiques, où des éléments différents remplissent la même fonction, assure stabilité et flexibilité, tout en favorisant l'innovation et l'exploration (Edelman et Gally, 2001; Seifert et al., 2016). L'exploration de nouvelles solutions motrices augmente cette dégénérescence, tandis que la stabilisation des solutions peut précéder un régime de métastabilité, où exploration et exploitation alternent (Komar et al., 2023; Seifert et al., 2015). L'analyse de clusters non supervisée permet d'identifier des patterns émergents sans a priori, surmontant les limites des méthodes classiques de variabilité (Chow et al., 2008). Les modèles de Markov dérivants (DMMs) quantifient les probabilités d'apparition de comportements et modélisent les transitions entre exploration et exploitation, comme observé dans l'acquisition de compétences motrices en nage (Komar et al., 2023) et en escalade (Kalligeris et al., 2024).

2 Les modèles de Markov dérivants

Les DMM sont une classe particulière de processus de Markov hétérogènes qui permettent à la matrice de transition de dériver (varier) le long de la séquence des observations.
Considérons un système aléatoire avec un espace d'états fini $A = \{1, 2, \ldots, s\}$ et soit $(\Omega, \mathcal{F}, P)$ un espace de probabilité. De plus, soient $(\Pi_0(u,v))_{u,v\in A}$ et $(\Pi_1(u,v))_{u,v\in A}$, des matrices de transition d'un modèle de Markov d'ordre 1 sur l'espace d'états A.

Definition 1 (Chaîne de Markov dérivante d'ordre 1 et de degré 1) *Une suite de variables aléatoires $X_0, \ldots, X_n$ dans un espace d'état A est appelée une chaîne de Markov dérivante d'ordre 1, de degré 1 et de longueur n entre $\Pi_0(u,v)$ et $\Pi_1(u,v)$, si la distribution des X_t, $t = 1, \ldots, n$, est définie par :*

$$\Pi_{\frac{t}{n}}(u,v) = P\left(X_t = v \middle| X_{t-1} = u\right) = \left(1 - \frac{t}{n}\right)\Pi_0(u,v) + \frac{t}{n}\Pi_1(u,v).$$

Definition 2 (Chaîne de Markov dérivante d'ordre k et de degré d)) *Soit $\Pi_{\frac{i}{d}}$, $i = 0, \ldots, d$, des matrices stochastiques sur A. Une suite de variables aléatoires $X_0, \ldots, X_n$ dans un espace d'état A est appelée une chaîne de Markov dérivante d'ordre k, de degré d et de longueur n entre $\Pi_{\frac{i}{d}}(u_1, \ldots, u_k, v)$, $i = 0, \ldots, d$,, si la distribution des X_t, $t = 1, \ldots, n$, est définie par :*

$$\Pi_{\frac{t}{n}}(u_1, \ldots, u_k, v) = P\left(X_t = v \middle| X_{t-1} = u_k, \ldots, X_{t-k} = u_1\right) = \sum_{i=0}^{d} p_i(t)\Pi_{\frac{i}{d}}(u_1, \ldots, u_k, v).$$

où $(u_1, \ldots, u_k) \in A^{k+1}$ *et* $p_i(t), i = \{0, 1, 2, \ldots\}$ *sont les polynômes de Lagrange de degré* d *tels que :*

$$\forall (i,j) \in \{0, \ldots, d\}^2, p_i\left(\frac{nj}{d}\right) = \mathbb{I}_{\{i=j\}}.$$

Avec les notations de la définition précédente, on a : (1) pour $t = \frac{ni}{d}$, on a $\Pi_{\frac{t}{n}} = \Pi_{\frac{i}{d}}$, $i = 0, \ldots, d$, et (2) $\forall t \in [0, n]$, $\sum_{v \in A} \Pi_{\frac{t}{n}}(u, v) = 1$.

Pour estimer $\Pi_{\frac{i}{d}}(u, v)$, $i = 0, \ldots, d$, on utilise une méthode point par point, que nous allons illustrer brièvement sur un DMM d'ordre k et de degré d. La fonction à minimiser est la suivante :

$$\sum_{t=1}^{n} \sum_{u \in A^k} \sum_{v \in A} \mathbb{I}_{\{X_{t-k} \ldots X_{t-1} = u\}} \left(\Pi_{\frac{t}{n}}(u, v) - \mathbb{I}_{\{X_t = v\}}\right)^2.$$

Pour déterminer les estimateurs $\widehat{\Pi}_{\frac{i}{d}}(u, v)$ de $\Pi_{\frac{i}{d}}(u, v)$, $\forall (u, v) \in A^k \times A$, nous devons résoudre le système suivant :

$$\left(\sum_{t=k}^{n} \mathbb{I}_{\{X_{t-k} \ldots X_{t-1} = u\}} p_i(t) p_j(t)\right)_{i,j \in [1,d]} \left(\widehat{\Pi}_{\frac{i}{d}}(u, v)\right)_{i \in [1,d]} = \left(\sum_{t=k}^{n} p_i(t) \mathbb{I}_{\{X_t = v, X_{t-k} \ldots X_{t-1} = u\}}\right)_{i \in [1,d]}.$$

Cette méthode d'estimation point par point, basée sur toutes les observations de la séquence, s'est avérée la plus efficace, en termes de vraisemblance. Pour plus d'informations sur cette méthode d'estimation ainsi que sur les propriétés des estimateurs obtenus, se référer à Vergne (2008) ; Barbu et Vergne(2019).

3 Applications des DMMs à la dynamique de l'acquisition des compétences

Dans une des premières applications des modèles de Markov dérivants (DMM) aux données comportementales, des questions clés sur les bénéfices de l'exploration pendant l'apprentissage ont été soulevées (Komar et al., 2023). L'application des DMM a permis de caractériser en profondeur la dynamique des patterns de mouvement émergents pendant l'apprentissage, en observant leur succession d'émergence et de disparition avec la pratique. Un modèle de degré 1 représente une évolution linéaire des transitions vers un pattern spécifique, sans exploration, associé à une disparition linéaire (coefficient négatif) ou à une émergence croissante (coefficient positif). En revanche, un modèle de degré supérieur à 1 est associé à des comportements exploratoires, montrant des fluctuations dans les probabilités d'apparition des patterns au cours de l'apprentissage, avec parfois un pattern dominant puis disparaissant. Ce modèle informe sur la nature exploratoire de ces patterns, fortement visités mais abandonnés, probablement jugés non optimaux par l'apprenant.

Pour illustrer cette approche, (Komar et al., 2023) ont identifié 11 clusters comportementaux à partir d'une analyse de clusters, représentant différents patterns de coordination bras-jambe pour la nage en brasse, qui ont été plus ou moins explorés par les apprenants, reflétant la quantité d'exploration motrice. Il a été observé que dans des conditions de haute vitesse, 12 différences significatives dans la quantité d'exploration des clusters ont été identifiées, alors qu'en basse vitesse, seules 2 différences significatives ont émergé, montrant une exploration quasi-égale de tous les clusters en basse vitesse. En revanche, dans des conditions de haute vitesse, les clusters 6 et 2 étaient plus explorés que d'autres, tandis que les clusters 8, 9 et 10 étaient moins explorés, suggérant que certains clusters étaient inaccessibles dans ces conditions. De plus, l'analyse de la dynamique individuelle d'exploration a montré que les apprenants ayant visité un nombre plus faible de clusters étaient plus susceptibles d'afficher une faible amplitude d'exploration pendant l'apprentissage. Par exemple, certains apprenants ont montré une stabilité élevée de quelques patterns de coordination, présents tout au long de la pratique. La figure 1 montre que le participant 7 a exploré uniquement quatre clusters en condition de faible vitesse, avec une probabilité élevée d'apparition du cluster 4 et secondairement du cluster 9, illustrant une bi-stabilité stable pendant les 16 sessions d'apprentissage. Cependant, exhiber un grand nombre de clusters n'était pas corrélé à une plus grande exploration ni à une amélioration de la performance, mais pouvait être associé à une faible stabilité, montrant des fluctuations aléatoires du comportement au fil des sessions, comme illustré à la figure 2 avec le participant 6 en condition de faible vitesse. Ces deux dynamiques, illustrées par les participants 6 et 7, ont conduit à des DMM de degré 1 et 3 très similaires, résultant en un faible niveau d'exploration (voir Figures 1 et 2).

Bien que certains apprenants (comme les participants 6 et 7) puissent afficher soit une grande stabilité d'un pattern dominant de coordination (c'est-à-dire un modèle de Markov dérivant montrant une probabilité élevée d'apparition), soit une instabilité de divers patterns de coordination (c'est-à-dire un modèle de Markov dérivant montrant une faible probabilité d'apparition de la plupart des clusters), la majorité des apprenants présente trois dynamiques au cours de l'apprentissage qui peuvent être identifiées par le DMM : i) l'apparition de nouveaux patterns de mouvement, ii) la disparition des patterns de mouvement initiaux/débutants, et iii) la dynamique exploratoire avec l'apparition/disparition successive des patterns de mouvement, correspondant à un régime métastable avec quelques patterns qui ne sont vraiment ni stables ni vraiment instables (voir Kelso (2012) pour le concept de métastabilité et Seifert et al. (2016) pour une application à l'acquisition de compétences). La figure 3 illustre ces trois dynamiques pour le participant 2 en condition de haute vitesse.

L'application des modèles de Markov dérivants (DMMs) permet d'analyser les patterns de mouvement émergents en identifiant ceux qui sont exploratoires ou stables au fil du temps. DMM aide à comprendre la dynamique d'apprentissage en identifiant des périodes clés : monostables (quand un seul pattern prédomine), instables (lorsqu'aucun pattern n'est plus probable que les autres) et métastables (lorsque certains patterns émergent mais de façon moins stable). Cela permet de mieux saisir l'évolution et la variabilité des comportements des apprenants pendant leur processus d'acquisition de compétences.

Dans une deuxième étude, l'application des DMMs a permis d'étudier l'acquisition des compétences visuo-motrices dans l'escalade, une tâche complexe nécessitant à la fois un

contrôle précis du mouvement actuel et une recherche anticipatoire des contraintes environnementales. L'analyse des temps de décalage du regard a permis d'identifier quatre clusters de comportement, allant de comportements exploratoires (moins efficaces) à des comportements plus exploitants (plus efficaces), où les apprenants ajustent et affinent leurs actions motrices. L'étude a montré que les apprenants alternent entre exploration et exploitation, ce qui reflète un régime intermittant, ou métastable, où de nouvelles solutions sont explorées tout en exploitant les solutions acquises.

Conclusion

L'application des modèles de Markov dérivants (DMMs) est particulièrement utile pour l'étude descriptive de la dynamique de l'acquisition des compétences (par exemple, par l'analyse de clusters) et permet de modéliser la probabilité d'apparition des différents modèles comportementaux visités, éventuellement stabilisés ou abandonnés par les apprenants. Ainsi, les DMMs ont permis de modéliser i) les différents parcours individuels d'apprentissage, et ii) la présence de modèles comportementaux coexistant, soit stabilisés (c'est-à-dire la multistabilité), soit simplement des tendances (c'est-à-dire la métastabilité).

Bibliographie

Barbu, V. S., Mavrogiannis, I., & Vergne, N. (2023). dsmmR: Estimation and simulation of drifting semi-markov models. `https://cran.r-project.org/web/packages/dsmmR/index.html`

Barbu, V. S., & Vergne, N. (2019). Reliability and survival analysis for drifting Markov models: Modeling and estimation. *Methodology and Computing in Applied Probability*, 21, 1407–1429.

Barbu, V. S., Vergne, N., Lothodé, C., & Seiller, A. (2021). drimmR: An R package for estimation, simulation, and reliability of drifting Markov models. `https://cran.r-project.org/web/packages/drimmR/index.html`

Button, C., Seifert, L., Chow, J., Araújo, D., & Davids, K. (2020). Dynamics of skill acquisition. *Journal of Motor Behavior*, 45(2), 123-130.

Chow, J. Y., Davids, K., Button, C., & Rein, R. (2008). Dynamics of movement patterning in learning a discrete multiarticular action. *Motor Control*, 12(3), 219-240.

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24), 13763-13768.

Haken, H. (1983). *Synergetics: An introduction: Nonequilibrium phase transition and self-organization in physics, chemistry, and biology*. Springer.

Kalligeris, E. N., Barbu, V., Hacques, G., Seifert, L., & Vergne, N. (2024). Unveiling the

persistent dynamics of visual-motor skill via drifting Markov modeling. *Nonlinear Dynamics, Psychology, and Life Sciences*, 28(4), 431-447.

Kelso, J. A. S. (2012). Multistability and metastability: Understanding dynamic coordination in the brain. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 367(1591), 906-918.

Komar, J., Seifert, L., Vergne, N., & Newell, K. M. (2023). Narrowing the coordination solution space during motor learning standardizes individual patterns of search strategy but diversifies learning rates. *Scientific Reports*, 13(1).

Seifert, L., Boulanger, J., Orth, D., & Davids, K. (2015). Environmental design shapes perceptual-motor exploration, learning, and transfer in climbing. *Frontiers in Psychology*, 6, 1819.

Seifert, L., Komar, J., Araújo, D., & Davids, K. (2016). Neurobiological degeneracy: A key property for functional adaptations of perception and action to constraints. *Neuroscience & Biobehavioral Reviews*, 69, 159-165.

Vergne, N. (2008). Drifting Markov models with polynomial drift and applications to DNA sequences. *Statistical Applications in Genetics and Molecular Biology*, 7(1). doi: doi:10.2202/1544-6115.1326

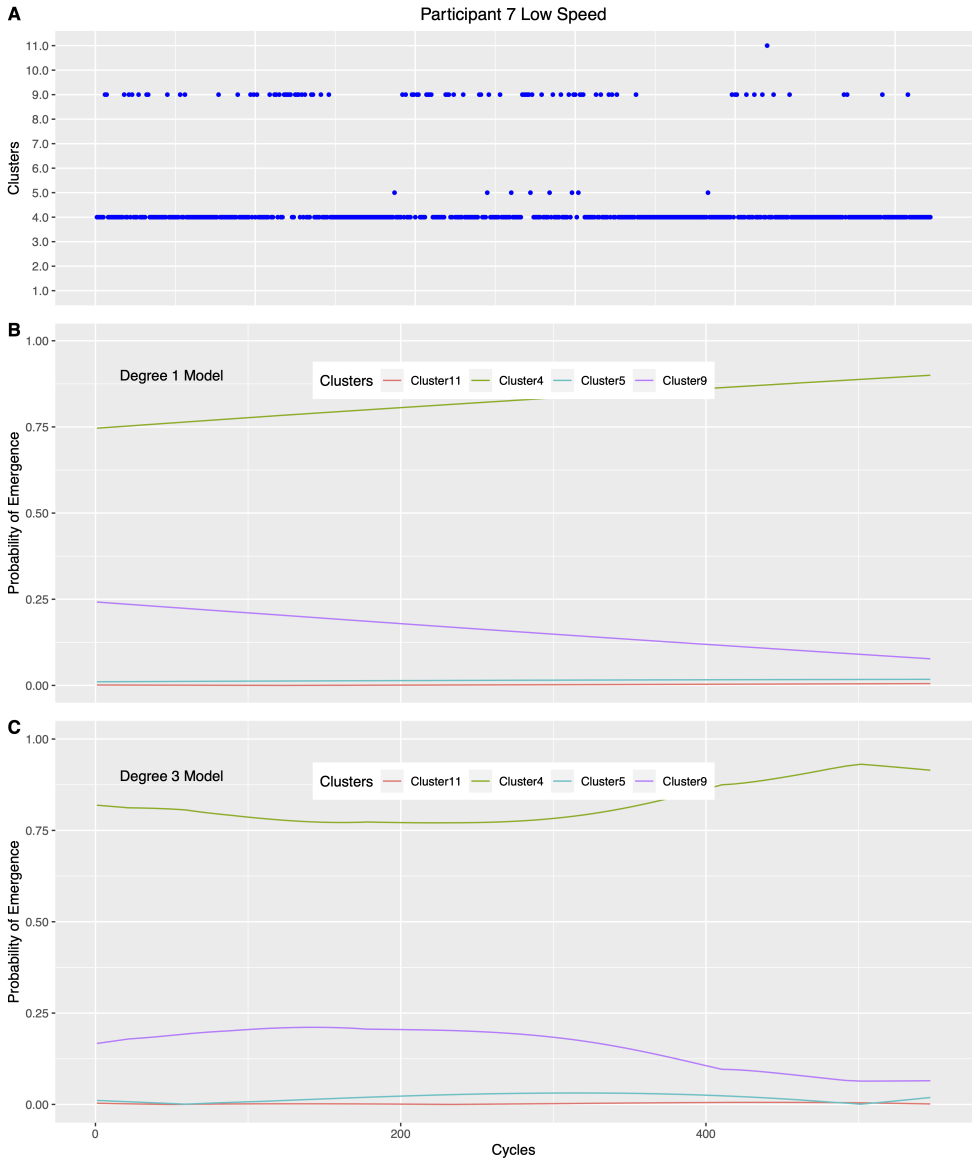


Figure 1: Dynamique individuelle de l'apprentissage pour le participant 7 en condition de faible vitesse montrant une bi-stabilité avec la dominance du cluster 4 et secondairement du cluster 9, tandis que la probabilité pour les clusters 5 et 11 est très faible. (A) Dynamique cycle par cycle (1 point représentant un cycle effectué) des patterns de coordination pendant l'ensemble du processus d'apprentissage (du cycle 1 au cycle 680). (B) Simulation des modèles de Markov dérivants entraînés sur les données de (A) montrant la probabilité d'apparition de chaque pattern pendant l'apprentissage avec une modélisation linéaire (c'est-à-dire de degré 1) des probabilités de transition entre les patterns. (C) Simulation des modèles de Markov dérivants entraînés sur les données de (A) montrant la probabilité d'apparition de chaque pattern pendant l'apprentissage avec une modélisation polynomiale de degré 3 des probabilités de transition entre les patterns. Pour chaque pattern, la quantité d'exploration est calculée comme la distance entre ces deux modèles dans (B) et (C).

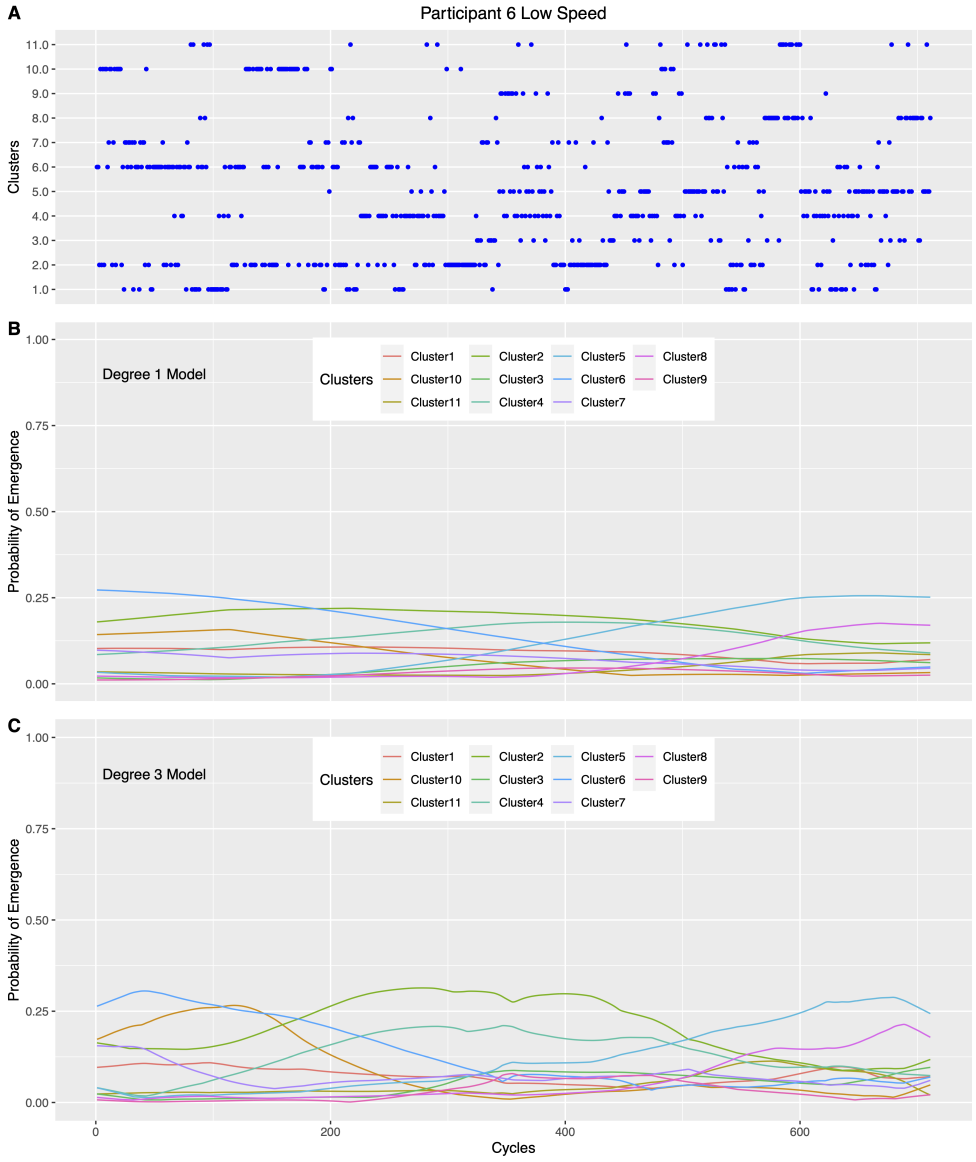


Figure 2: Dynamique individuelle de l'apprentissage et de l'exploration pour le participant 6 en condition de faible vitesse montrant une instabilité avec une faible variation de la probabilité d'apparition du cluster tout au long du processus d'apprentissage. (A) Dynamique cycle par cycle (1 point représentant un cycle effectué) des patterns de coordination pendant l'ensemble du processus d'apprentissage (du cycle 1 au cycle 680). (B) Simulation des modèles de Markov dérivants entraînés sur les données de (A) montrant la probabilité d'apparition de chaque pattern pendant l'apprentissage avec une modélisation linéaire (c'est-à-dire de degré 1) des probabilités de transition entre les patterns. (C) Simulation des modèles de Markov dérivants entraînés sur les données de (A) montrant la probabilité d'apparition de chaque pattern pendant l'apprentissage avec une modélisation polynomiale de degré 3 des probabilités de transition entre les patterns. Pour chaque pattern, la quantité d'exploration est calculée comme la distance entre ces deux modèles dans (B) et (C).

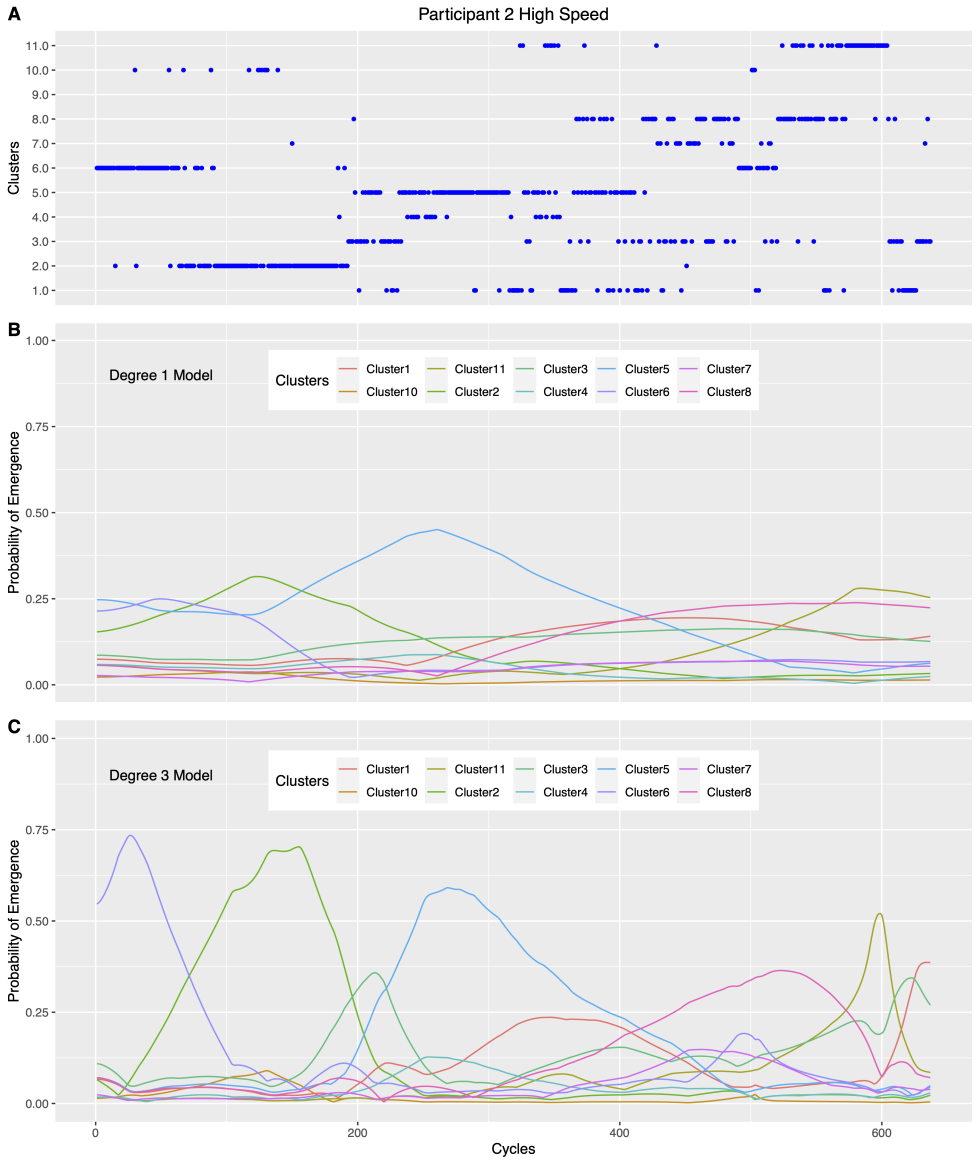


Figure 3: Dynamique individuelle d'apprentissage et d'exploration pour le participant 2 en condition de haute vitesse montrant un régime métastable dans les dynamiques d'apprentissage (autour des cycles 400 à 450). (A) Dynamique cycle par cycle (1 point représentant un cycle effectué) des patterns de coordination pendant l'ensemble du processus d'apprentissage (du cycle 1 au cycle 680). (B) Simulation des modèles de Markov dérivants entraînés sur les données de (A) montrant la probabilité d'apparition de chaque pattern pendant l'apprentissage avec une modélisation linéaire (c'est-à-dire de degré 1) des probabilités de transition entre les patterns. (C). Simulation des modèles de Markov dérivants entraînés sur les données de (A) montrant la probabilité d'apparition de chaque pattern pendant l'apprentissage avec une modélisation polynomiale de degré 3 des probabilités de transition entre les patterns. Pour chaque pattern, la quantité d'exploration est calculée comme la distance entre ces deux modèles dans (B) et (C).

PSOLxG : Une Nouvelle Métrique pour Décoder la Performance des Gardiens et Attaquants au Football

Rémi Verronneau[1] & Vianney Perchet[2]

[1] *FootoVision et CREST,ENSAE ;* remi.verronneau@footovision.fr
[2] *CREST,ENSAE et CRITEO AI LAB ;* vianney.perchet@ensae.fr

Résumé Nous présentons une nouvelle méthodologie pour affiner l'analyse de la probabilité de marquer au football, à travers la métrique **PSOLxG** (Post Shot On The Line expected Goal). Contrairement aux approches d'expected Goals (xG) qui se concentrent sur la probabilité de but avant la frappe, le PSOLxG s'intéresse à la trajectoire, la vitesse et la localisation exacte du ballon lorsqu'il se dirige vers le but. En modélisant ces facteurs supplémentaires, il devient possible de distinguer plus finement la contribution du gardien (difficulté de son arrêt) de celle de l'attaquant (qualité de frappe). Notre modèle s'appuie sur un *Explainable Boosting Machine* (EBM), une forme avancée de modèle additif général (GAM) enrichi de quelques termes d'interaction. Cette structure combinant performance prédictive et interprétabilité permet d'analyser comment chaque variable (par exemple la vitesse ou la hauteur de frappe) influence localement la probabilité de but. Grâce à une calibration rigoureuse, les probabilités prédites correspondent aux fréquences empiriques, évitant ainsi les biais de sur- ou sous-estimation. Nous construisons ensuite des indicateurs pour évaluer les performances individuelles. Pour les gardiens, l'agrégation des PSOLxG et des buts encaissés aboutit à un **GKR** (Goalkeeper Ratio) permettant de comparer des portiers ayant fait face à différents volumes de tirs. Pour les attaquants, nous introduisons le **STR** (Striker Ratio), qui compare la valeur xG initiale d'un tir à la probabilité post-frappe (PSOLxG). Ces indicateurs révèlent la capacité d'un joueur à bonifier ou à détériorer une situation donnée. Enfin, une méthode de **projection** permet d'anticiper la performance d'un joueur changeant de championnat : pour chaque tir réalisé dans un pays, on recherche des tirs similaires dans la ligue cible et on compare les probabilités de but correspondantes. Cette démarche offre un éclairage précieux pour le recrutement et les analystes de club.

Mots-clés. Sciences du football, métriques d'évaluation, modèles additifs, apprentissage, interprétabilité.

Abstract We introduce a new methodology to refine soccer scoring probability analysis through the **PSOLxG** metric (Post Shot On The Line expected Goal). Unlike traditional expected Goals (xG) approaches that focus on the likelihood of scoring before the strike, PSOLxG considers the ball's trajectory, speed, and exact location as it travels toward the goal. By modeling these extra factors, it becomes possible to better isolate the goalkeeper's contribution (save quality) from the striker's contribution (shot quality). Our model relies on an *Explainable Boosting Machine* (EBM), an advanced form of Generalized Additive Model (GAM) enriched with a few interaction terms. This structure—combining predictive performance and interpretability—enables us to examine how each variable (e.g., shot speed or height) locally influences the scoring probability. With careful calibration, predicted probabilities match empirical frequencies, thus avoiding systematic over- or underestimation bias. We then propose individual performance metrics. For goalkeepers, aggregating PSOLxG and conceded goals yields the **GKR** (Goalkeeper Rating), which allows comparisons among keepers facing different shooting volumes. For attackers, we introduce the **STR** (Striker Ratio), which compares the initial xG value of a shot to its post-strike probability (PSOLxG). These indices highlight a player's ability to "enhance" or degrade a given opportunity. Lastly, a **projection** method estimates how a player's performance might translate when moving between leagues: for each shot taken in one country, we search for the most similar shots in the target league and compare their associated scoring probabilities. This process offers valuable insights for player recruitment and strategic planning.

Keywords. Soccer analytics, goalkeeper and strikers metrics, additive models, interpretable ML.

1 Introduction

En football, un seul but peut changer l'issue d'un match entier. Contrairement à d'autres sports où les points sont fréquents, la rareté et l'importance d'un but en football rendent chaque occasion de marquer cruciale. Les attaquants brillent en transformant des chances improbables en buts, tandis que les gardiens de but démontrent leur talent en arrêtant des tirs apparemment imparables.

Le concept des buts attendus (**expected Goal ou xG**) est devenu central dans l'analyse du football (Gelman and Hill, 2006; Decroos et al., 2019). Les modèles traditionnels d'xG estiment la probabilité de marquer *avant le tir*, en utilisant des facteurs comme la distance, l'angle ou la partie du corps utilisée. Cependant, ces modèles négligent souvent les particularités du tir, telles que la trajectoire du ballon, sa vitesse ou sa position finale.

Pour répondre à ces limites, nous introduisons le **PSOLxG**, qui estime la probabilité de but à partir de variables telle que la position de la balle au départ et à l'arrivée ou encore la vitesse du ballon, une fois celui-ci connu comme *on-target*. Le PSOLxG est donc particulièrement utile pour évaluer les gardiens de but puisqu'on peut le comparer avec le résultat réel de l'action. Nous proposons aussi un indicateur pour évaluer la qualité de finisseur de l'attaquant en distinguant clairement la probabilité initiale (xG) et la probabilité post-frappe (PSOLxG). Afin de conserver une grande lisibilité tout en capturant des phénomènes non linéaires et les *interactions* entre variables (ex: vitesse *vs* hauteur de frappe), nous utilisons un *Explainable Boosting Machine* (EBM) Lou et al. (2013); Caruana et al. (2015). Ce modèle, lorsqu'il est bien calibré, offre à la fois une prédiction fiable et une *interprétabilité locale* (décomposition de la contribution de chaque variable pour un tir donné).

Cette contribution a donc pour but d'exploiter la métrique PSOLxG que nous avons créée pour comparer et évaluer des gardiens de buts et des attaquants. Dans une optique de changement de circonstance, nous présentons également une méthode pour projeter les performances d'un attaquant d'un championnat à un autre.

2 Cadre Méthodologique et Mathématique

Formalisation du problème Soit $y_i \in \{0, 1\}$ l'issue d'un tir (1 si but). Chaque tir est décrit par un vecteur x_i de 7 variables, incluant `startX`, `startY` (position initiale du tireur), `gkX`, `gkY` (position du gardien), et `endY`, `endZ` (coordonnées finales du ballon projetées orthogonalement sur la ligne de but $x = 0$), plus `speedKmh`. En donnant la coordonnée `endY` du tir dans le cas où le gardien stoppe la balle, il existe un risque que l'on donne une partie de la réponse (but ou pas but). Pour palier à cela, nous projetons le tir sur la ligne de but ($x = 0$), là où l'on enregistre les coordonnées finales en cas de but. Nous faisons l'hypothèse simplificatrice d'une trajectoire linéaire en projection 2D jusqu'à $x = 0$.

Les GAM sont intéressants car ils autorisent des effets non linéaires pour chaque variable, tout en conservant une structure additive. Pour un échantillon (x_i, y_i), on modélise la log-odds :

$$\log\Big(\tfrac{p_i}{1-p_i}\Big) \;=\; \sum_j f_j(x_{ij}),$$

où $p_i = P(y_i = 1 \mid x_i)$ et f_j des fonctions univariées (ex: splines).

Les *EBMs* étendent les GAMs en y ajoutant *des termes d'interaction par paires.* Pour chaque tir x_i, les EBMs modélisent la log-odds comme :

$$g\left(P(y_i = 1 \mid x_i)\right) = \underbrace{\sum_j f_j(x_{ij})}_{\text{termes univariés}} + \underbrace{\sum_{j<k} f_{jk}(x_{ij}, x_{ik})}_{\text{interactions par paires}},$$

avec $g(\cdot)$ la fonction logit utilisée en régression logistique $g(p) = \log\left(\frac{p}{1-p}\right)$, $f_j(\cdot)$ la fonction univariée (comme dans un GAM), apprise via un boosting de faibles arbres de décision (quelques niveaux de profondeur) et $f_{jk}(\cdot,\cdot)$ une fonction bivariée pour représenter l'interaction entre deux variables (ex: `endY` et `speedKmh`).

En limitant les interactions à des dimensions faibles (univariées ou bivariées), les EBMs restent visualisables et donc interprétables. On peut observer séparément l'effet de chaque variable ou de chaque couple de variables sur la probabilité de but, ce qui s'avère particulièrement adapté pour les analyses sportives.

La calibration du modèle est cruciale ici. C'est-à-dire que ses probabilités prédites doivent correspondre aux fréquences empiriques de buts. Une mauvaise calibration peut déformer l'évaluation de la performance d'un gardien (par exemple en exagérant ses capacités d'arrêt) ou d'un attaquant (par exemple en sous-estimant sa finition clinique). Il est donc indispensable de veiller à ce que les probabilités du modèle concordent avec les observations. Un modèle mal calibré pourrait prédire 70% de chances de marquer pour certains tirs, alors que seulement 55% aboutissent réellement. Une telle erreur systématique entraînerait des évaluations biaisées des gardiens ou des stratégies de tir.

Optimisation Bayésienne pour le réglage d'Hyperparamètres Lorsqu'on entraîne un EBM, des hyperparamètres tels que *le taux d'apprentissage (learning rate)*, *le nombre d'itérations de boosting* et *la complexité d'interaction* influencent la précision et l'interprétabilité. Il est souvent délicat de trouver un compromis entre performance prédictive suffisamment élevée et une bonne calibration. Pour résoudre ce problème, nous adoptons une stratégie d'**optimisation bayésienne** (Snoek et al., 2012) où l'on considère la fonction objectif (somme pondérée de log-loss et d'ECE) comme inconnue puis on modélise cette fonction via un processus gaussien, mis à jour à chaque test d'une nouvelle configuration d'hyperparamètres. Enfin on choisit la prochaine configuration grâce à une fonction d'acquisition. On peut aussi réaliser un tuning classique (minimisation du log-loss seul) et améliorer la calibration *a posteriori* via un ajustement post-calibration comme la régression isotone proposée par Zadrozny and Elkan (2002).

3 Résultats Empiriques et Interprétation

Classification et calibration Nous entrainons notre modèle sur 5 000 tirs de plusieurs championnats professionnels européens de différents niveaux (année 2023-2024). Le but est d'obtenir des résultats pour le niveau moyen d'un joueur professionnel. Le modèle distingue globalement les buts des non-buts avec une précision solide ($\approx 84\%$). Un ROC AUC de 0.9046 traduit un bon pouvoir discriminant. On note qu'il identifie plus facilement les *non-buts* (F1 à 0.90) que les *buts* (F1 à 0.70), phénomène courant en raison d'un déséquilibre de classes (moins de buts). Néanmoins, un rappel de 0.65 pour la classe "But" reste satisfaisant pour capturer la majorité des vrais buts. L'ECE finale est autour de 0,02, indiquant que les probabilités produites correspondent globalement aux fréquences réelles de but. Dans la version longue, on compare les résultats à plusieurs autres modèles de ML utilisés pour la classification. Les modèles à base d'arbres affichent de meilleurs résultats. Intuitivement,la construction des terrains de football est naturellement adaptée aux modèles arborescents, qui peuvent diviser l'espace en régions régions distinctes où les probabilités de marquer des buts diffèrent.

Exemple d'Interprétation Locale Les EBMs permettent de décomposer la log-odds d'un tir en contributions de chaque variable (et de chaque interaction). Les Figures 1 et 2 (issues d'un tableau de bord Footovision) illustrent un exemple de tir prédit à 74,8% de chance de marquer.

Nous donnons ici les valeurs de quelques variables de ce tir et nous proposons une interprétation :

- `endY : 3.04` (près du poteau) et `speedKmh : 97.04` (vitesse élevée) contribuent *positivement* à la

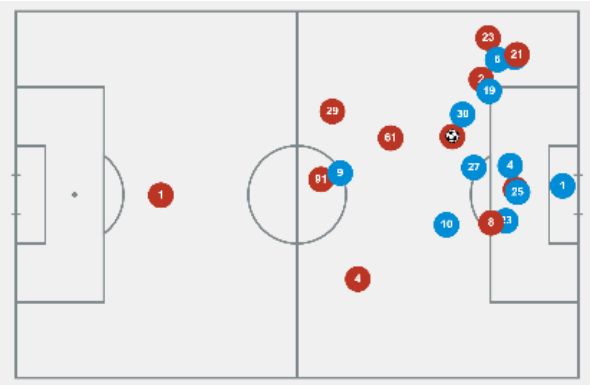


Figure 1: Vue 2D du tir (sur la plateforme Footovision). *L'équipe rouge attaque de gauche à droite.*

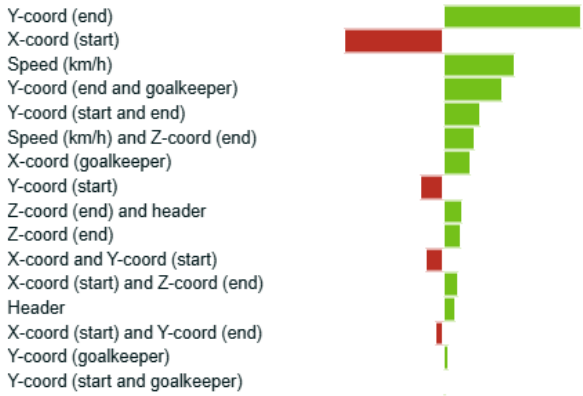


Figure 2: Explication locale de la prédiction pour ce tir. *Probabilité de but : 0,748*

- `startX : 21` (distance plus grande) pénalise légèrement la probabilité de marquer, sans pour autant compenser les autres facteurs plus favorables.
- L'interaction $(\texttt{startY}, \texttt{endY})$ renforce la log-odds, révélant qu'un tir partant d'un certain angle vers un point opposé proche du poteau est plus dangereux.

De telles explications locales sont précieuses pour conforter l'expertise métier (*"Les tirs en lucarne sont difficiles à arrêter"*) et découvrir des synergies subtiles (*"Une hauteur de balle particulière devient problématique si la vitesse est dans la fourchette particulière"*).

4 Création de métriques d'évaluation pour les gardiens et les attaquants

Évaluation des gardiens sur leur ligne En additionnant respectivement les xG et les PSOLxG de tous les tirs qu'un gardien a subis, on peut construire la métrique *GKR* (**G**oal**K**eeper **R**atio) définie comme suit:

$$\text{GKR} = \frac{\sum_i \text{PSOLxG}_i - \text{Total_Conceded_Goals}}{\text{Total_Conceded_Shot}}.$$

Une valeur de GKR supérieure à 0 indique que le gardien a encaissé moins de buts que prévu (surperformance) ; une valeur négative indique le contraire. Cette formule, normalisée par le nombre de tirs, permet de comparer des gardiens ayant disputé un nombre différent de matchs. Elle ne favorise pas non plus un gardien ayant encaissé moins de buts "par chance" d'avoir une bonne défense. Par exemple, si deux gardiens ont fait face à 50 tirs chacun, mais que le premier a encaissé 5 buts alors que le PSOLxG total est de 10, et que le second en a encaissé 10 pour 15 de PSOLxG, ils auront la même GKR. Nous appliquons le modèle aux gardiens anglais sur la première moitié de la saison 2024–2025 (jusqu'en décembre 2024). Nous cumulons les PSOLxG de tous les tirs reçus, comparons aux buts encaissés, puis calculons le ratio ci-dessus. Fait marquant : le gardien de Nottingham Forest, Selz, se classe troisième selon ce ratio, alors que son équipe est étonnamment dans le top 4 à mi-saison — devant des clubs prestigieux comme Manchester City et United. Il n'a concédé que 14 buts, alors que notre modèle en prédisait 17, évitant ainsi trois buts supplémentaires. À l'inverse, le gardien portugais José Sá en a concédé 3,5 de plus que prévu par le modèle, sur 39 tirs cadrés analysés, et son équipe est proche de la relégation.

Évaluer la Capacité de Finition d'un attaquant: Convertir les xG en PSOLxG Pour mesurer la capacité d'un attaquant à améliorer une situation de tir (estimée par xG) en un PSOLxG plus élevé

différence cumulée. L'idée est que la technique, le choix de frappe et le placement d'un attaquant peuvent modifier la probabilité de but *une fois la balle frappée*. Les tentatives hors cadre reçoivent un PSOLxG de zéro. On introduit un indicateur simple qu'on note STR (**ST**riker **R**atio) :

$$STR^1 = \frac{\sum_i \mathbf{PSOLxG}_i \; - \; \sum_i \mathbf{xG}_i}{\text{Total_Shot}}.$$

Un ratio positif suggère que les tirs de l'attaquant sont *plus dangereux* que ce que prévoyait le xG initial. Un ratio négatif laisse entendre que l'attaquant *sous-performe* ses attentes, possiblement en raison d'une finition moins bonne.

Cette métrique met aussi en avant des joueurs qui n'ont pas marqué de buts comme le joueur de couloir de Brighton, Enciso, qui aurait "mérité"de marquer 2.5 buts (son total de PSOLxG), pour seulement 1.5 xG initial sur 41 tentatives. On remarque cependant qu'un joueur comme Salah (Liverpool) obtient seulement 7 PSOLxG pour 5 xG, alors qu'il a tout de même inscrit 10 buts. Au football, marquer un but reste l'objectif premier, même si on ne l'atteint pas toujours via des frappes parfaites. Pour tenir compte de cela, nous définissons une variante STR^2 :

$$STR^2 = \frac{\sum_i \max(\mathbf{PSOLxG}_i, \mathbf{Goal}_i) - \sum_i \mathbf{xG}_i}{\text{Total_Shot}},$$

où l'on prend le maximum entre PSOLxG et l'indicateur de but (1 si but, 0 sinon). Ainsi, un but marqué (même avec un PSOLxG faible) ne pénalise pas injustement un attaquant. D'autre part, un ratio positif indique toujours un dépassement des attentes initiales (xG) ; un ratio négatif, une sous-performance.

Attaquant	**Somme Max**	**Somme xG**	**Tot Buts**	**Tot Tirs**	STR^2
Mbeumo (Bre)	7.9	1.9	7	20	+0.3
Wood (Not)	10.47	3.51	9	24	+0.29
...	...	...	...	...	...
Tavernier (Bou)	1.49	1.96	1	18	-0.02
Rutter (Bri)	1.2	1.75	1	17	-0.03

Avec cette nouvelle mesure, l'attaquant Wood de Nottingham Forest est dans le top 2, son équipe étant deuxième du championnat à mi-saison et suscitant la surprise. Mbeumo, le buteur de Brentford, se hisse en tête du classement, avec un total "PSOLxG ou but" bien supérieur à son nombre de buts effectifs, témoignant de frappes de grande qualité face à des gardiens eux aussi performants.

Projection dans un autre championnat Lorsqu'un club souhaite recruter un joueur d'un autre championnat, il ne sait pas à l'avance comment le joueur va s'adapter dans un nouvel environnement. L'idée ici est de projeter les performances de frappe d'un joueur du championnat portugais dans une ligue supérieure, par exemple la Premier League anglaise. Pour cela, nous prenons chaque tir enregistré au Portugal et nous recherchons les tirs les plus similaires en Angleterre. Nous utilisons la méthode des *K plus proches voisins* (k-NN) introduit par Cover and Hart (1967). Nous avons fixé $\mathbf{k = 3}$ après hyperparamétrisation. Nous comparons la moyenne des *PSOLxG* des 3 tirs anglais les plus proches du *PSOLxG* réel du tir portugais correspondant en se basant sur les résultats du k-NN.

Pour chaque tir portugais, on calcule la moyenne du *PSOLxG* des 3 tirs anglais les plus similaires, et on compare à la valeur réelle du PSOLxG. On définit alors la mesure $TRAM$ (**TRA**nsfer **M**etric) pour chaque joueur sur l'ensemble de ses tirs i:

$$TRAM = \frac{\sum_i \big(\mathbf{PSOLxG}_i - \mathbf{PSOLxG}_{\text{KNN_England_i}}\big)}{\text{Total_Shots}},$$

où pour un tir i, $\mathbf{PSOLxG}_{\text{KNN_England_i}}$ est la moyenne des PSOLxG des 3 tirs similaires à ce tir au

c'est qu'il fait mieux, en moyenne, qu'un collègue en Angleterre sur un tir similaire. Le joueur du club de Braga (top 4), Bruma, semble être le plus propice à un changement de circonstance et obtient un TRAM de 0.01. En d'autres termes, il obtient un *PSOLxG* supérieur de 0,01 en moyenne, par tir, à celui de ses homologues anglais dans des situations similaires, sur un échantillon d'environ 25 tirs par joueur. On note qu'en réalisant l'opération sur une saison complète, un échantillon plus grand renforcerait la robustesse des analyses.

5 Conclusion

Cette étude met en avant de nouveaux indicateurs, tels que le **GKR** pour les gardiens et le **STR** pour les attaquants, afin de quantifier objectivement leurs performances respectives. Ces KPI (Key Performance Indicator) exploitent la métrique **PSOLxG**, dérivée d'un modèle *Explainable Boosting Machine*, comme un outil d'évaluation précis et interprétable. Le PSOLxG n'est pas une finalité mais un levier permettant de mieux contextualiser les performances des joueurs en intégrant des données fines comme la trajectoire et la vitesse du ballon. Ces contributions offrent une base solide pour une analyse comparative des joueurs, tout en ouvrant la voie à des applications dans le recrutement et les stratégies de transfert. Cependant, la performance du modèle dépend fortement de la qualité des données, souvent affectée par des informations spatiotemporelles incomplètes ou des protocoles de collecte variables selon les ligues. Les trajectoires non linéaires et des facteurs contextuels clés, tels que la pression défensive ou les tactiques d'équipe, ne sont pas toujours pleinement pris en compte. De plus, l'analyse de transfert peut introduire des biais, car les styles de jeu, la physicalité et le niveau de compétition varient entre les championnats. Ainsi, le PSOLxG doit être considéré comme un outil complémentaire à l'expertise humaine. Les travaux futurs pourraient enrichir la métrique en intégrant des données plus précises sur les trajectoires ou des caractéristiques avancées capturant mieux les subtilités du football d'élite.

References

Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*, pages 623–631, 2013.

R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, pages 1721–1730, 2015.

T. Cover and P. Hart. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

T. Decroos, J. Van Haaren, and J. Davis. Actions Speak Louder than Goals: Valuing Player Actions in Soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 1851–1861, 2019.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.

B. Zadrozny and C. Elkan. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 694–699, 2002.

Statistique appliquée

Vers l'automatisation de la synthèse de catalyseurs grâce à la spectroscopie Raman *in situ* en temps réel couplée à l'intelligence artificielle

Clara Dubois [1,2,3], Joëlle Thuriot-Roukos [1], Christophe Dujardin [1*] , Sophie Dabo-Niang [2] & Sébastien Legendre [3]

[1] *Université de Lille, CNRS, Centrale Lille, Univ. Artois, UMR 8181, UCCS - Unité de Catalyse et Chimie du Solide, 59000 Lille, France*
**christophe.dujardin@centralelille.fr*
[2] *Université de Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, INRIA-MODAL, 59000 Lille, France*
[3] *HORIBA France SAS, 455 Avenue Eugène Avinée, 59120 Loos, France*

Résumé. L'optimisation et l'automatisation de la synthèse chimique constituent un enjeu majeur pour améliorer l'efficacité des procédés. Dans ce projet, nous développons une méthodologie statistique combinant apprentissage automatique et spectroscopie afin de faciliter l'analyse des réactions chimiques et l'optimisation des conditions expérimentales. Les données spectroscopiques, riches en informations mais souvent complexes, nécessitent une réduction de dimension pour en extraire les tendances principales. L'analyse en composantes principales (ACP) permet de simplifier cette complexité en maximisant la variance expliquée tout en offrant une visualisation intuitive des schémas cachés dans les spectres. Par ailleurs, des méthodes de clustering sont employées pour identifier des regroupements naturels dans les spectres, mettant en évidence des comportements chimiques distincts, notamment des étapes de réactions. En complément, des approches d'apprentissage supervisé sont utilisées pour prédire des propriétés chimiques à partir des spectres acquis. Différents modèles de régression sont évalués pour estimer certains paramètres expérimentaux, tandis que des algorithmes de classification permettent d'attribuer automatiquement un spectre à une catégorie donnée. Ces outils offrent un cadre robuste pour automatiser l'analyse des données et guider la synthèse chimique de manière plus efficace. Les résultats obtenus indiquent que ces méthodes facilitent l'analyse des données spectroscopiques et apportent des éléments utiles pour l'optimisation des processus de synthèse. Cette approche pourrait contribuer à une automatisation plus efficace en laboratoire et à une meilleure exploitation des données expérimentales grâce à l'intelligence artificielle.

Mots-clés. Apprentissage Statistique/automatique, spectroscopie, optimisation de synthèse

Abstract. Optimizing and automating chemical synthesis is a major challenge for improving process efficiency. In this project, we are developing a statistical methodology combining machine learning and spectroscopy to facilitate the analysis of chemical reactions and the optimization of experimental conditions. Spectroscopic data is rich in information but often complex, requiring a reduction in dimension to extract the main trends. Principal component analysis (PCA) helps to simplify this complexity by maximizing the variance explained, while

offering an intuitive visualization of the patterns hidden in the spectra. In addition, clustering methods are used to identify natural groupings in the spectra, highlighting distinct chemical behaviors, in particular reaction stages. In addition, supervised learning approaches are used to predict chemical properties from the spectra acquired. Different regression models are evaluated to accurately estimate certain experimental parameters, while classification algorithms are used to automatically assign a spectrum to a given category. These tools provide a robust framework for automating data analysis and guiding chemical synthesis more efficiently. The results obtained indicate that these methods facilitate the analysis of spectroscopic data and provide useful elements for optimizing synthesis processes. This approach could contribute to more efficient automation in the laboratory and better exploitation of experimental data using artificial intelligence.

Keywords. Statistical/Machine learning, Spectroscopy, Synthesis optimization

1 Introduction

Avec la demande croissante de produits manufacturés dans de nombreux secteurs, il devient crucial de repenser certains modes de production afin de minimiser leur impact sur l'environnement. D'après Zybert et al, près de 90% des procédés chimiques industriels impliquent l'utilisation de la catalyse[1]. La catalyse a donc un rôle essentiel à jouer pour répondre aux défis sociétaux en réduisant l'empreinte des procédés sur la planète grâce à une recherche qui s'inscrit dans le cadre du développement durable. Dans ce contexte, l'optimisation des procédés de synthèse des catalyseurs, associée au développement de technologies plus durables et efficaces, devient essentielle.
Cette recherche se concentre sur le développement d'une nouvelle méthode de synthèse des composés d'intérêts, notamment des catalyseurs, en cherchant à optimiser leur fabrication pour qu'elle soit à la fois plus respectueuse de l'environnement et plus performante. Le projet se concentre sur l'intégration des outils de synthèse avec ceux de caractérisation pour assurer leur compatibilité et leur bon fonctionnement. Ce processus opérationnel permettra d'adapter une synthèse en temps réel en utilisant une démarche raisonnée basée sur l'intelligence artificielle.

L'évolution des techniques analytiques dans le domaine de la chimie a permis une meilleure compréhension des systèmes complexes et une amélioration des processus industriels. Nous avons choisi la spectroscopie Raman pour l'identification des composés présents dans le milieu réactionnel car elle présente de nombreux avantages. La spectroscopie Raman est une technique d'analyse non destructive qui mesure la diffusion inélastique de la lumière lorsqu'elle interagit avec les molécules d'un échantillon. Cette technique consiste à détecter les changements de fréquence du laser après interaction avec les vibrations moléculaires induisant un changement du tenseur de polarisabilité[2]. Elle fournit des informations sur la composition du milieu réactionnel directement en solution, sans préparation complexe de l'échantillon. Cependant, le traitement des données massives générées par cette technique, notamment les spectres, reste un défi majeur. L'intelligence artificielle (IA) offre une approche innovante pour relever ce défi, en permettant d'extraire des informations précieuses à partir

des spectres et de prédire les propriétés chimiques associées[3]. Ce document présente une méthodologie combinant à la fois IA supervisée et non supervisée pour la prédiction de propriétés chimiques et la classification des spectres, ainsi que pour l'exploration non supervisée visant à différencier les spectres dans des systèmes inconnus.

2 Processus expérimental et collecte de données

Dans cette étude, un réacteur innovant a été développé pour permettre un suivi en temps réel de réactions chimiques avec plusieurs précurseurs et différentes conditions de réaction. L'adaptation implique la mise en place d'un système de surveillance en temps réel des paramètres clés tels que le pH et la température. Notre approche est basée sur l'utilisation de la spectroscopie Raman pour collecter des données spectrales tout au long du processus de réaction. La figure 1 montre une série de spectres Raman, correspondant à un précurseur chimique à base de molybdène où un acide à été ajouté progressivement, permettant ainsi d'explorer un ensemble de la plage de pH, allant de pH=8 à pH=1. Les données spectrales acquises sont ensuite utilisées pour former des modèles d'apprentissage automatique, qui sont adaptés pour modéliser et améliorer les performances de notre système. Ces modèles devraient permettre de gérer les conditions opératoires à partir des spectres Raman référencés. L'idée est que le modèle d'apprentissage automatique soit capable de prédire l'évolution du processus et de prendre les décisions adéquates pour assurer le bon déroulement de la synthèse.

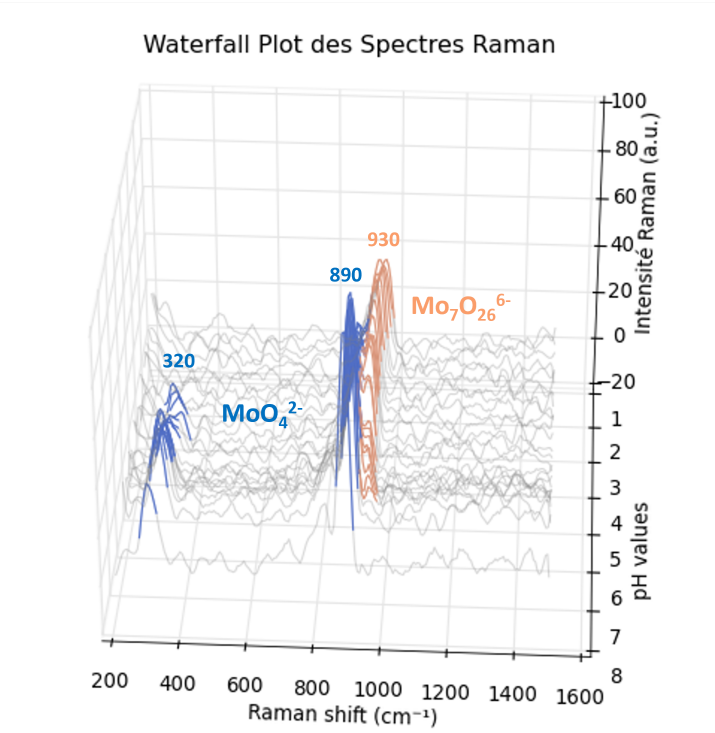


FIGURE 1 – Spectres Raman d'une solution à base de molybdène

Deux approches ont été étudiées, l'apprentissage non supervisé, tel que l'analyse en

composantes principales (ACP) pour explorer et réduire la dimensionnalité du spectre et le clustering, et les méthodes d'apprentissage supervisé telles que la régression et la classification qui ont démontré leur efficacité dans l'optimisation des processus de synthèse[4].

3 Apprentissage non-supervisé : réduction de dimension et clustering

L'apprentissage non supervisé, constitue une méthode usuelle d'explororation de données sans informations préalables sur leurs catégories ou organisations implicites[5]. Parmi les méthodes courantes, on retrouve le clustering par k-Means et les techniques de réduction de dimension, telle que l'analyse en composante principale (ACP), qui permet de réduire la complexité des données et d'en faciliter la visualisation et l'interprétation. Dans le contexte chimique, ces approches peuvent être utilisées pour identifier et regrouper des spectres similaires, même en l'absence de connaissances préalables sur la nature du système étudié ou les mécanismes réactionnels impliqués. Ces spectres peuvent provenir de diverses techniques d'analyse, telles que la spectroscopie infrarouge (IR), la spectroscopie Raman ou la résonance magnétique nucléaire (RMN), qui fournissent chacune des informations spécifiques sur les caractéristiques physiques ou chimiques des échantillons.
L'analyse en composantes principales (ACP) de données fonctionnelles[6] a permis de réduire la dimensionnalité des spectres, les deux premières composantes expliquant 75% de la variance totale dans le cas du précurseur de molybdène. La projection des observations sur le plan factoriel met en évidence la linéarité du processus chimique réalisé. Ces résultats ont ensuite été approfondis par une méthode de clustering k-means, permettant d'identifier 3 clusters distincts. Par exemple, le cluster rouge regroupe des spectres que l'on pourrait interpréter comme représentant les "débuts de la réaction". C'est à dire qu'ils correspondent aux spectres caractéristiques de l'espèce présente au début de la réaction chimique. Le cluster vert va plutôt correspondre à un état de transition, où le pH commence à devenir de plus en plus acide, donc la première espèce commence à disparaître et une nouvelle espèce fait son apparition simultanément. Enfin, le cluster bleu fera référence à la fin de la réaction, avec des pH très acides, où il n'y aura plus que les pics correspondant à l'espèce finale.

4 Apprentissage supervisé : Régression et Classification

L'apprentissage supervisée repose sur l'utilisation de données étiquetées (représentés par une variable réponse Y comme le pH) pour entraîner des modèles de régression ou de classification. Dans le cadre de cette étude, deux objectifs principaux sont poursuivis : la prédiction de propriétés chimiques ainsi que la classification des spectres.

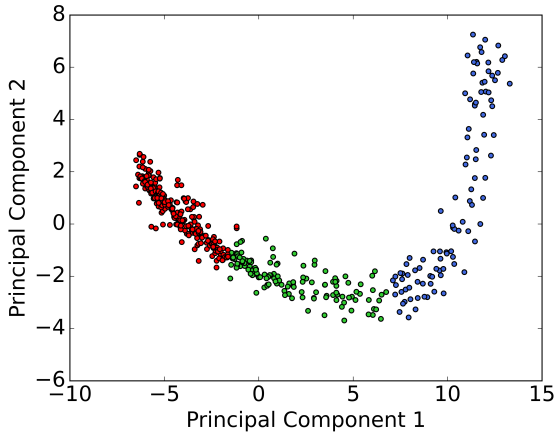


FIGURE 2 – Clustering k-means sur les spectres Raman d'une solution de molybdène à différents pH

4.1 Prédiction de propriétés chimiques

Dans cette étude, plusieurs méthodes de régression, telles que la régression linéaire ou celle des k-plus proches voisins, ont été mises en œuvre pour prédire des propriétés chimiques, notamment le pH, du système chimique. Les données d'entrées (variables explicatives X_i) sont les spectres Raman (les 10 premières composantes de l'analyse en composantes principales). Tous les modèles de régression ont été entraînés sur 80% du jeu de données, et les tests ont été effectués avec les 20% restants.

Les performances de cinq modèles de régression ont été évaluées à l'aide de métriques comme R^2, montrant que tous les modèles offrent de bons résultats. Cependant, le modèle de régression par Bagging se démarque légèrement avec une précision de 96%, comme le montre la figure 3.

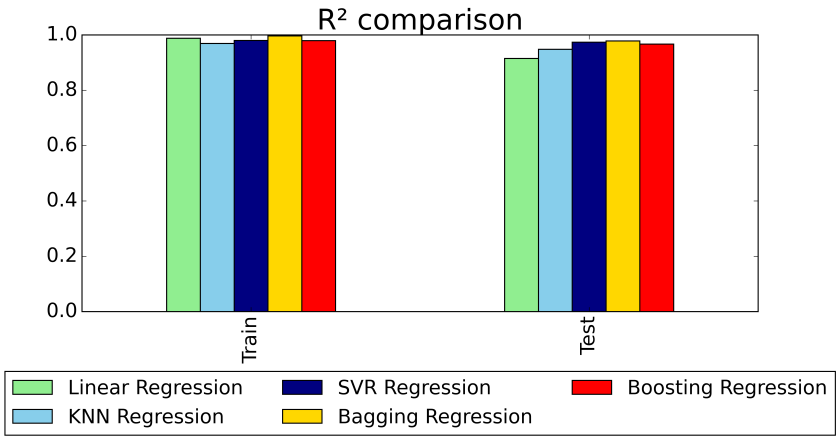


FIGURE 3 – Comparaison de différents modèle de régression pour la prédiction du pH

La figure 4 montre que le modèle de régression par échantillonnage (Bagging) donne des prédiction précise des valeurs de pH à travers l'ensemble de la plage de pH où a lieu la

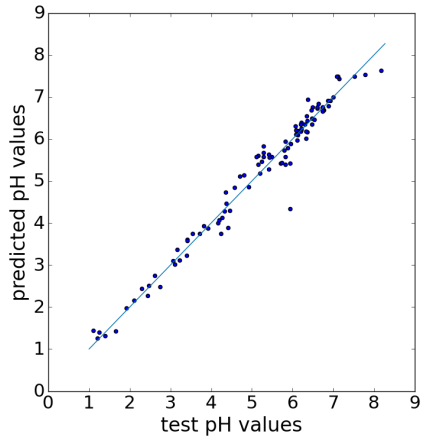


FIGURE 4 – Diagramme de corrélation pour la prédiction du pH de solutions de Na_2MoO_4 avec le modèle de régression par Bagging

réaction.

4.2 Classification des spectres

L'objectif est ici de catégoriser les spectres en fonction de l'espèce chimique représentée. Plusieurs méthodes de classification, telles que les machines à vecteurs de support (SVM), k-plus proches voisins (KNN) ou analyse discriminante quadratique (QDA) ont été utilisées pour classer les spectres Raman acquis tout au long de notre réaction chimique.

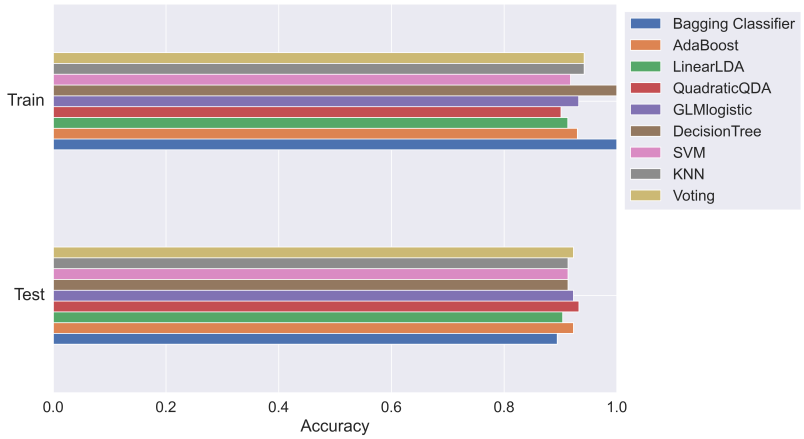


FIGURE 5 – Comparaison de 9 modèles de classification de spectres Raman de Na_2MoO_4

Les performances des modèles ont été comparés dans les figures 5-6 à l'aide de métriques comme le taux de spectres correctement classés ou la matrice de confusion. Parmi ces approches, le modèle de classification logistique s'est distingué, atteignant une précision de 92%. Comme le montre la figure 6, pour un spectre appartenant à la classe n°1, le modèle de classification logistique a correctement prédit la catégorie avec une probabilité de 95%. Ces

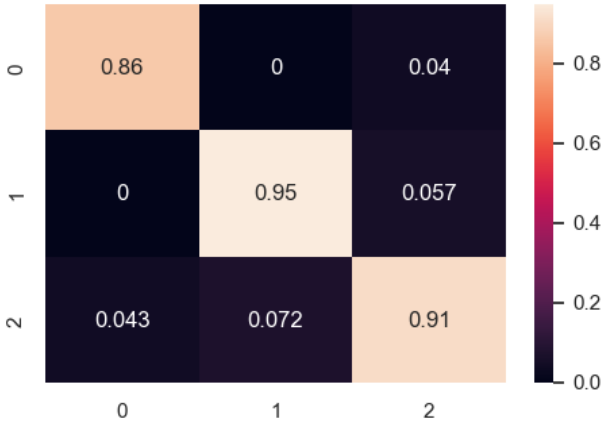


FIGURE 6 – Matrice de confusion associée au modèle de classification logistique

résultats démontrent l'efficacité du modèle de classification pour un système chimique simple impliquant une réaction avec un précurseur unique.

5 Perspectives

Cette étude d'un système chimique relativement simple (un précurseur unique et monophasique) a également été réalisé avec un deuxième précurseur, le tungstate de potassium (K_2WO_4). L'apprentissage non supervisé s'est montré pertinente pour explorer la structure des données et expliquer le processus de la réaction. Les méthodes d'apprentissages supervisés, quant à elles, se montrent intéressantes pour prédire et catégoriser certaines propriétés chimiques spécifiques au système chimique étudié. La première perspective est de créer une base de données qui engloberait tous les précurseurs, afin d'entraîner les modèles à différencier les spectres Raman en fonction de la position des pics présents, car un spectre Raman est considéré comme une empreinte digitale d'une molécule. De plus, l'idée future est d'étendre cette approche à des systèmes chimiques plus complexes, par exemple avec une phase de précipitation durant la réaction. L'objectif est de se rapprocher autant que possible des étapes d'une synthèse conventionnelle, qui inclut très fréquemment une phase de précipitation. Pour ce faire, nous prévoyons d'ajouter au moins un paramètre d'analyse supplémentaire à fournir pour l'entraînement des modèles afin de mieux détecter le moment de la précipitation.

Références

[1]M. ZYBERT, "Applied Catalysis in Chemical Industry : Synthesis, Catalyst Design, and Evaluation", Catalysts (Basel, Switzerland) **13**, 10.3390/catal13030607 (2023).

[2]P. VANDENABEELE, *Practical Raman Spectroscopy an Introduction* (Wiley, Chichester, West Sussex, U.K, 2013).

[3]A. Angulo, L. Yang, E. S. Aydil et M. A. Modestino, "Machine learning enhanced spectroscopic analysis : towards autonomous chemical mixture characterization for rapid process optimization", Digital Discovery **1**, 35-44 (2022).

[4]F. Mekki-Berrada, Z. Ren, T. Huang, W. K. Wong, F. Zheng, J. Xie, I. P. S. Tian, S. Jayavelu, Z. Mahfoud, D. Bash, K. Hippalgaonkar, S. Khan, T. Buonassisi, Q. Li et X. Wang, "Two-Step Machine Learning Enables Optimized Nanoparticle Synthesis", npj computational materials **7**, 1-10 (2021).

[5]T. Hastie, R. Tibshirani et J. Friedman, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*, Springer, 2017.

[6]J. Ramsay et B. W. Silverman, "Functional Data Analysis (Springer Series in Statistics)", (1997).

Des modes d'évolution culturelle contrastés : Langues kra-dai et technologies de tissage

Emma KOPP [1] & Robin J.Ryder [2] & Chris Buckley [3] & Guillaume Jacques [4] & Thomas Pellard [5]

[1] *CEREMADE, Université Paris-Dauphine, France, kopp@ceremade.dauphine.fr*
[2] *Department of Statistics, Imperial College, UK, ryder@ceremade.dauphine.fr*
[3] *Tracing Patterns Foundation, USA, chrisbuckley888@hotmail.com*
[4] *Institut National des Langues et Civilisations Orientales, CNRS, France, rgyalrongskad@gmail.com*
[5] *Institut National des Langues et Civilisations Orientales, CNRS, France, thomas.pellard@gmail.com*

Résumé. Nous étudions et comparons l'évolution de deux aspects de la culture, les langues et les technologies de tissage, chez les Kra-Dai (Tai-Kadai) du sud-ouest de la Chine et du sud-est de l'Asie, en utilisant des méthodes bayésiennes de Markov-Chain Monte Carlo pour découvrir des phylogénies. Les résultats montrent que les langues et les métiers à tisser ont évolué de manière apparentée mais différente, et apportent de nouvelles informations sur la diaspora des locuteurs Kra-Dai à travers l'Asie du Sud-Est. Nous avons constaté que les langues et les métiers à tisser utilisés par les locuteurs Hlai de Hainan sont des groupes marginaux dans les phylogénies linguistiques et de métiers à tisser, et que les métiers à tisser utilisés par les locuteurs de langues étroitement apparentées ont tendance à appartenir à des types similaires. Cependant, nous avons également constaté des différences à un niveau profond, tant dans les détails de l'évolution des métiers à tisser et des langues que dans leurs schémas généraux de changement.

Mots-clés. Inférence Bayésienne, Phylolinguistique, Language, Évolution culturelle, Statistiques computationelles.

Abstract. We study and compare the evolution of two aspects of culture, languages and weaving technologies, among the Kra-Dai (Tai-Kadai) of southwest China and southeast Asia, using Bayesian Markov-Chain Monte Carlo methods to discover discover phylogenies. The results show that languages and looms evolved in related but different ways, and provide new insights into the diaspora of Kra-Dai speakers across Southeast Asia. We found that the languages and looms used by Hlai speakers from Hainan are marginal groups in linguistic and loom phylogenies, and that looms used by speakers of closely related languages tend to belong to similar types. However, we also found differences at a deep level, both in the details of loom and language evolution and in their general patterns of change, and we discuss possible reasons for these differences.

Keywords. Bayesian Inference, Phylolinguistic , Language, Cultural Evolution, Computational Statistics.

1 Introduction

Les langues et les techniques de tissage sont toutes deux des traditions complexes transmises de génération en génération avec des modifications. Au fil du temps, ces modifications s'accumulent et les traditions évoluent : ce fait permet aux chercheurs d'étudier leur histoire en utilisant des modèles arborescents. Ces dernières années, les méthodes phylogénétiques, en particulier celles utilisant des approches bayésiennes, sont devenues une technique courante. Elles permettent d'inférer la topologie et les longeurs de branche d'un arbre d'évolution d'un groupe d'organismes et de répondre à diverses questions. Entre outre, il est courant d'inférer l'âge de l'ancêtre commun le plus récent (MRCA) au groupe d'organisme étudié. Pour reconstituer l'histoire des langues (Gray and Atkinson, Gray et al., Chang et al., Kolipakam et al., Sagart et al., Heggarty et al.). Les méthodes phylogénétiques ont également été appliquées à d'autres aspects de la culture (Mace and Holden), y compris à la culture matérielle et aux technologies. Les méthodes utilisées incluent les méthodes basées sur la distance (Jordan and Shennan, Saslis-Lagoudakis et al.) et la parcimonie maximale (Buchanan and Collard,Tehrani et al.,O'Brien et al., Le Bomin et al.), et, dans une moindre mesure, les méthodes bayésiennes

(Matthews et al., Buckley and Boudot [a] Learmouth et al.). Plusieurs études ont examiné les similarités et différences entre l'évolution des artefacts et celle des langues (Jordan and O'Neill, Jordan and Shennan, Passmore et al.). Ces observations soulèvent la question plus large de savoir si, et dans quelle mesure, les langues et d'autres caractéristiques culturelles évoluent de manière similaire.

Les Kra-Dai sont l'un des nombreux groupes répartis dans les provinces du sud de la Chine et dans plusieurs pays d'Asie du Sud-Est continentale, notamment le Vietnam, le Laos, la Thaïlande et le Myanmar, ainsi que dans la province de l'Assam, dans le nord-est de l'Inde. Le tissage constitue un aspect important de la culture des Kra-Dai, tant pour des raisons pratiques (fabrication d'objets utiles tels que vêtements, couvertures ou moustiquaires) que pour l'expression de l'identité personnelle et du statut social. Ces pratiques se transmettent majoritairement de génération en génération.

Une phylogénie est un arbre daté sur lequel on définit un processus évolutif stochastique caractérisé par des transitions entre états. Sa structure est définie par une Chaîne de Markov, c'est-à-dire que les états futurs ne dépendent que de l'état présent et soient indépendants des états passés. Soit un arbre enraciné (avec racine r), avec est une collection de feuilles V (de taille L) et un ensemble des arêtes E. Chaque noeud a un âge, défini comme sa distance de la racine. On note $Y \in \mathbb{R}^{LK}$ les observations binaires aux feuilles où K est le nombre de sites observés et L le noombre de taxons observés. On émet l'hypothèse que pour chaque site k, les vecteurs $\left(Y_1^k, \ldots, Y_L^k\right)$ évoluent indépendamment.

Pour modéliser une phylogénie, on choisit une loi à priori sur l'arbre. Celle-ci régit la topologies et la longueurs des branches. Nous considérerons un modèle de Naissance-Mort Gernhard [2008] pour la loi à priori sur l'arbre. Sur cet arbre évolutif, un événement de diversification peut se produire avec une probabilité τ (taux de naissance). Chaque lignée est sujette à l'extinction avec un taux de mort μ . Les langues sont échantillonnées selon une proportion α, ce qui reflète la conscience du modèle qu'il ne dispose que d'un sous-ensemble de l'ensemble complet des langues, sans prendre en compte les langues éteintes. On choisit ensuite un processus d'évolution des données le long de l'arbre définit par la matrice de transition Q, dont l'élément q_{ij} représente le taux exponentiel de passage de l'état i à l'état j. Dans cet article, nous considérons deux modèles : le modèle de chaîne de Markov en temps continu (CTMC), où il n'y a que deux état 0 et 1 et le modèle covarion à deux états, où l'on considère 4 états tels que chacun des deux états (0 et 1) acceptent un mode rapide et un mode lent. La matrice Q inclue un taux d'évolution temporel de chaque branche. Il convertit le nombre de mutations d'une langue en temps. Par défaut, chaque substitution dans l'arbre se produit à taux λ, le nombre moyen de substitutions par site par unité de temps. Pour rendre les modèles plus réalistes, il est fréquent d'autoriser de l'hétérogénité parmis les sites ou parmis les lignées de l'arbre. Dans le premier cas, on considère que chaque site k évolue selon ses propres paramètre Q^k. Dans le second, chaque branche e de l'arbre est caractérisé par un paramètre λ_e, le nombre moyen de substitutions par site par unité de temps sur la branche e.

Ayant établi un modèle, nous devons maintenant identifier l'arbre qui s'aligne le mieux avec les données disponibles. Pour cela, nous utilisons BEAST2 (Suchard et al. [2018]), un logiciel puissant pour implémenter des algorithmes MCMC robustes. Il génère un ensemble d'arbres échantillonnés à partir de la distribution à posteriori : pour chaque analyse, ces arbres ont été résumés en calculant un arbre de consensus selon la règle de la majorité, qui montre les nœuds présents dans au moins 50 % des arbres échantillonnés. Pour évaluer la qualité relative des ajustements de modèle, nous avons calculé la vraisemblance marginale (Marginal Likelihood, ML) de chaque modèle.

Dans cette étude, nous comparons les schémas d'évolution des langues et des techniques de tissage chez les Kra-Dai d'Asie orientale et d'Asie du Sud-Est. Nous avons choisi les techniques de tissage parce que (comme les langues) leurs caractéristiques essentielles sont transmises de manière conservatrice, généralement de parent (ou de proche parent) à enfant au sein des sociétés rurales. Elles sont riches en termes de complexité et de variation (comme les langues) et offrent (en principe du moins) un champ d'innovation considérable. Comme pour le vocabulaire linguistique de base, certaines technologies du métier à tisser ont des durées de vie considérables, comparables à celles des langues Buckley and Boudot [b].

2 Données

Les linguistes encodent les mots en classe de cognats. Les cognats sont des mots apparentés qui descendent d'un même ancêtre linguistique commun. Par exemple, les sens français « doigt »et espagnol « dedos »partagent une origine commune, ce qui les place dans la même classe de cognats pour ce sens précis. En revanche, les mots français « doigt »et anglais « finger »ne sont pas considérés comme des cognats directs, bien qu'ils désignent le même objet. Cela suggère que « doigt »et « dedos »partagent probablement un ancêtre commun, tandis que « dedos »et « finger »ne le font pas. Ainsi pour le sens « doig »et les langues français espagnol et anglais, on identifie deux classe de cognat : doigt, dedos et finger. Notre jeu de données sur les langues Kra-Dai est basé sur celui de Tao et al., une base de données de vocabulaire de base comprenant des formes lexicales pour 90 concepts dans 100 langues, annotées pour la cognation, avec un total de 647 ensembles de classes de cognats, encodés sous forme de matrice de traits binaires présence/absence de la classe de cognat.

Notre jeu de données sur les technologies de tissage Kra-Dai (métiers à tisser) est basé sur celui de Buckley and Boudot [a]. Il comprend 29 métiers à tisser et de 216 sites. Comme pour les langues, elles enregistrent la présence ou l'absence d'un ensemble de caractères, dans ce cas, les sites représentent des d'attributs fonctionnels tels que la barre de chaîne, la barre de tissu, le batteur de trame, etc. Les langues et les métiers à tisser diffèrent sur un point essentiel : alors qu'il est courant de modéliser les formes linguistiques cognates comme variant indépendamment, cette hypothèse peut être remise en question pour certains traits des métiers à tisser qui dépendent d'autres éléments. Cette interdépendance est une caractéristique générale des technologies, où les composants se construisent sur la présence d'autres composants. Pour évaluer si l'interdépendance des caractères influence les inférences phylogénétiques, nous avons attribué des « niveaux »aux caractères des métiers à tisser. Le niveau 1 comprend les caractères totalement indépendants les uns des autres. Le niveau 2 regroupe les caractères qui sont des sous-catégories de ceux du niveau 1, et ainsi de suite jusqu'au niveau 4.

3 Résultats

La convergence a été vérifiée pour tous les modèles testés. Dans l'ensemble, nous avons constaté que les arbres de consensus obtenus pour les métiers à tisser et les langues variaient peu entre les modèles. Pour des raisons limite de rendu, nous utiliserons uniqument la Figure 1 qui compare les arbres phylogénétiques de consensus pour les langues et les métiers à tisser (pour des raisons de clarté, seuls les taxons présents dans les deux ensembles de données sont affichés). Les modèles choisis pour les deux études, après comparaisons de modèle à l'aide du calcul de la vraissemblance marginale sont le modèle binaire covarion avec hétérogénéité des taux (dans les branches et les sites).

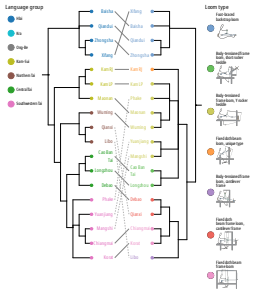


FIGURE 1 – Comparaison des arbres de consensus à règle de majorité calculés pour les langues (à gauche) et les métiers à tisser (à droite), restreints aux taxons présents dans les deux ensembles de données.

3.1 Dépendance entre les traits et les taux de changement des traits

Pour évaluer l'influence potentielle de l'interdépendance des traits sur nos résultats, nous avons mené trois analyses distinctes des caractéristiques des métiers à tisser, en attribuant des pondérations différentes à chaque niveau. Dans la première analyse, nous avons inclus uniquement les traits de niveau 1, considérés comme entièrement indépendants les uns des autres. L'arbre de consensus obtenu selon la règle de la majorité révèle trois clades principaux : les métiers à sangle dorsale simples avec appui pour les pieds, les métiers à cadre tendus par le corps (incluant des types avec lisses fixées à de longs leviers en Y ou à des leviers courts) et les métiers à cadre avec rouleau d'enroulement fixe, ces derniers englobant les types les plus complexes. Les branches récentes de cet arbre sont peu résolues, ce qui est attendu, car nous avons exclu les données des niveaux 2, 3 et 4, qui décrivent l'élaboration évolutive des caractéristiques de base des métiers à tisser.

3.2 Taux d'évolution pour les métiers à tisser

Le codage des caractéristiques des métiers à tisser par niveaux a permis de réaliser une analyse distincte en attribuant quatre taux de mutation différents, un pour chaque niveau de dépendance. L'arbre phylogénétique consensuel obtenu présente la même topologie que dans les analyses précédentes, ce qui confirme la robustesse de cette structure face aux détails du modèle évolutif. Cependant, des différences significatives dans les taux de changement ont été observées entre les quatre niveaux 2. Les traits de niveau 1, supposés indépendants les uns des autres et correspondant aux caractéristiques de base des métiers, présentent le taux de mutation le plus bas. Les traits de niveau 2 ont un taux plus élevé, et ceux de niveau 3 le plus élevé. Pour les traits de niveau 4, les données sont moins claires et le taux de mutation présente une plus grande variabilité, probablement en raison du petit nombre de caractères (13 pour le niveau 4, contre 102 pour le niveau 2).

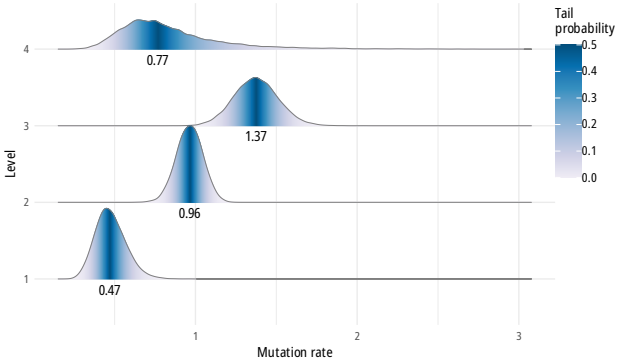


FIGURE 2 – Fonction de densité de probabilité et valeurs médianes de la distribution des taux de mutation dans les métiers à tisser, selon le niveau de trait

3.3 Résultats des analyses phylogénétiques des métiers à tisser et des langues

Malgré de nombreuses différences de détail, les deux arbres présentent une structure globalement similaire. Les lignes pleines indiquent les correspondances compatibles avec les deux phylogénies, tandis que les lignes en pointillés signalent des informations conflictuelles. Les métiers à tisser et les langues les plus divergents par rapport au reste des peuples Kra-Dai sont ceux des Hlai (bleu moyen, près du sommet du diagramme). La partie centrale des arbres est occupée par les locuteurs restés près des terres d'origine des Tai dans les provinces de Guizhou et de Guangxi en Chine, qui conservent des formes de métiers à tisser plus anciennes avec cadres en L. Les locuteurs SWT, aux métiers à tisser plus innovants, apparaissent dans le tiers inférieur de la figure, répartis de la frontière sino-vietnamienne jusqu'au Laos et à la Thaïlande.

La Figure 3 présentent les distributions d'âges pour les ancêtres communs les plus récents (MRCA) des langues Kra-Dai, Kam-Tai et Tai-Yay, selon l'analyse avec horloge relaxée et taux hétérogène. L'analyse avec

horloge stricte donne un âge racine considérablement plus ancien et irréaliste (âge médian de 8966 ans contre 5333 ans avec l'horloge relaxée), probablement en raison de la plus grande longueur de branche du groupe Kra. Cela peut s'expliquer par plusieurs facteurs : les langues Kra se sont divergées tôt du reste du groupe Kra-Dai, et de nombreuses langues de ce groupe ont subi un nombre significatif de changements phonétiques, ce qui pourrait indiquer une évolution lexicale plus rapide.

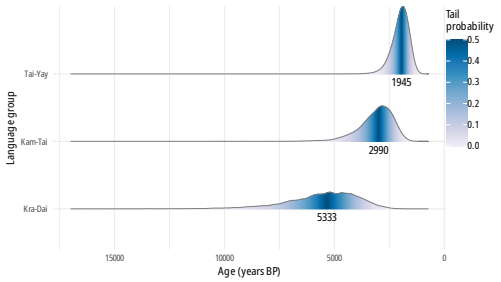


FIGURE 3 – Fonction de densité de probabilité et valeurs médianes de la distribution des âges (en année BP) pour les ancêtres communs les plus récents (temps de divergence) des langues Kra-Dai, Kam-Tai et Tai-Yay (modèle covarion binaire, taux hétérogènes).

4 Discussion

Dans cette partie, nous examinons les similitudes entre les phylogénies des métiers à tisser et des langues.

Les phylogénies des langues et des métiers à tisser partagent plusieurs caractéristiques communes dans leurs topologies. Les deux montrent une division précoce entre les Hlai de Hainan et le reste de la famille Kra-Dai, et leurs structures imbriquées reflètent les emplacements géographiques des taxons et les voies migratoires présumées. La comparaison de la phylogénie des métiers à tisser dans la Figure 1 montre que les différents clades sont principalement définis par des caractéristiques structurelles des cadres, et que celles-ci semblent s'être développées de manière progressive, avec des caractéristiques plus récentes ajoutées aux caractéristiques plus anciennes plutôt que de les remplacer. Cette propriété émerge naturellement de l'analyse phylogénétique et est indépendante de notre affectation des niveaux aux caractéristiques des métiers à tisser. On peut également observer que les métiers à tisser avec des cadres plus simples (principalement en forme de L, avec les caractéristiques de Niveau 1 et Niveau 2) se trouvent près de l'origine présumée des peuples parlant le Tai dans le sud de la Chine, tandis que les métiers à tisser avec des cadres plus complexes et des caractéristiques de Niveau 3 (barres supérieures et cadres en forme de boîte) se trouvent principalement dans le nord du Vietnam, au Laos et en Thaïlande, parmi les locuteurs des langues SWT.

4.1 Mesurer la similitude entre les métiers à tisser et les langues

Nous avons également évalué quantitativement le degré de similarité entre les arbres phylogénétiques des métiers à tisser et des langues. Nous souhaitons répondre à deux questions : premièrement, les phylogénies des métiers à tisser et des langues sont-elles compatibles ? Deuxièmement, est-il raisonnable de supposer que les métiers à tisser et les langues ont évolué selon exactement le même arbre phylogénétique ?

Diverses méthodologies ont été utilisées dans la littérature pour traiter des questions similaires. Brown et al. et Passmore et al. calculent des matrices de distances entre feuilles dans plusieurs ensembles de données, puis effectuent des tests de corrélation entre ces matrices.

Pour la première question, nous avons élagué les données et les arbres pour ne conserver que les feuilles présentes dans les deux ensembles de données, et réalisé deux tests statistiques. D'abord, nous avons comparé l'ensemble des distances représentées par les arbres consensuels. Pour chaque paire de feuilles, nous avons calculé la distance patristique dans les arbres consensuels, de manière similaire à Brown et al. : cela correspond

à la longueur du chemin entre deux feuilles, via leur ancêtre commun le plus récent. Cette procédure a produit une matrice de distances entre langues, et une matrice de distances entre métiers à tisser. Nous avons effectué cette analyse en moyennant sur 1000 arbres échantillonnés à partir de la distribution postérieure de chaque jeu de données, en normalisant toutes les longueurs de branches par la profondeur des arbres pour éviter de donner un poids excessif aux arbres plus profonds. Pour chaque analyse et chaque paire de feuilles, nous avons calculé la distance moyenne sur les arbres. En comparant les deux matrices de distances moyennes, nous avons à nouveau trouvé des preuves significatives de corrélation (test de Mantel, $z = 292{,}94$, $p < 0{,}001$).

Ensuite, nous avons mesuré si l'arbre des métiers à tisser permet d'expliquer les données linguistiques, et inversement. Pour ce faire, nous avons réalisé une analyse en composantes principales (PCA) sur les données binaires aux feuilles. Nous avons ensuite calculé le K de Blomberg Blomberg et al., qui mesure la force du signal phylogénétique dans les données aux extrémités de l'arbre, en projetant la première composante des données des métiers à tisser sur les arbres linguistiques (modèle covarion binaire, taux hétérogènes) et celle des données linguistiques sur les arbres des métiers à tisser (mêmes modèles, tous niveaux, sans pondération), en moyennant à chaque fois sur l'ensemble des arbres de l'échantillon postérieur. En projetant les données des métiers à tisser sur les arbres linguistiques postérieurs, nous obtenons $K = 5{,}87$ ($p = 0{,}001$) ; en projetant les données linguistiques sur les arbres des métiers à tisser, nous obtenons $K = 2{,}22$ ($p = 0{,}001$). Le K de Blomberg représente le rapport entre la variance observée expliquée par la phylogénie et la variance attendue selon un modèle aléatoire ; des valeurs supérieures à 1, comme ici ($K = 5{,}87$ et $K = 2{,}22$), indiquent un signal phylogénétique fort. La valeur $K = 5{,}87$ suggère que la phylogénie linguistique explique une part importante de la variation dans les données sur les métiers à tisser.

Pour la seconde question, nous avons comparé différents modèles à l'aide de facteurs de Bayes. Nous avons calculé la vraisemblance marginale pour deux scénarios : (A) les métiers à tisser et les langues ont évolué selon la même phylogénie, et (B) les métiers à tisser et les langues ont évolué selon deux phylogénies distinctes. Pour cela, nous avons restreint notre analyse aux taxons présents dans les deux ensembles de données et utilisé le même modèle pour les deux ensembles (covarion binaire, horloge stricte, taux homogène).

Le scénario A suppose que les traits des métiers à tisser et des langues ont évolué selon un seul et même arbre phylogénétique. Nous avons donc fusionné les matrices de caractères des métiers à tisser et des langues, et lancé l'analyse. Le scénario B, en revanche, suppose que les traits des métiers à tisser et des langues ont évolué selon des phylogénies différentes. Nous avons donc analysé séparément les deux ensembles de données, et additionné les vraisemblances marginales des deux analyses. Nous avons obtenu une vraisemblance logarithmique marginale de -3229 (écart type 5,02) pour le scénario A, et de -3093 (écart type 4,18) pour le scénario B. Cela correspond à un facteur de Bayes logarithmique de 135 en faveur du modèle B. En suivant l'échelle standard de Kass and Raftery, il s'agit donc d'une preuve décisive en faveur du scénario B : nous concluons que, malgré les similarités évoquées plus haut, les phylogénies des langues et des métiers à tisser sont distinctes.

5 Conclusion

Cette étude comparée de l'évolution des langues et des métiers à tisser chez les peuples Kra-Dai montre que ces deux dimensions culturelles suivent des trajectoires parallèles mais distinctes. Les métiers à tisser présentent une évolution modulaire et ponctuée, tandis que les langues évoluent de manière plus régulière. Les résultats confirment une origine commune suivie de migrations et de diversifications, notamment l'expansion récente des langues SWT. L'analyse statistique révèle une forte corrélation entre les arbres phylogénétiques des langues et des métiers à tisser, mais les facteurs de Bayes indiquent de façon décisive que ces deux ensembles de traits ont évolué selon des phylogénies distinctes. Ces différences soulignent des dynamiques évolutives propres à chaque domaine culturel, liées à l'intensité des pressions sélectives.

Références

S. P. Blomberg, T. Garland, and A. R. Ives. Testing for phylogenetic signal in comparative data. 57(4) :717–745. doi : 10.1111/j.0014-3820.2003.tb00285.x.

S. Brown, P. E. Savage, A. M.-S. Ko, M. Stoneking, Y.-C. Ko, J.-H. Loo, and J. A. Trejaut. Correlations in the population structure of music, genes and language. 281(1774). doi : 10.1098/rspb.2013.2072.

B. Buchanan and M. Collard. Investigating the peopling of north america through cladistic analyses of early paleoindian projectile points. 26(3) :366–393. doi : 10.1016/j.jaa.2007.02.005.

C. D. Buckley and E. Boudot. The evolution of an ancient technology. 4(5) :170208, a. doi : 10.1098/rsos.170208.

C. D. Buckley and E. Boudot. The evolution of an ancient technology. 4(5) :170208, b. doi : 10.1098/rsos.170208.

W. Chang, C. Cathcart, D. Hall, and A. Garrett. Ancestry-constrained phylogenetic analysis supports the indo-european steppe hypothesis. 91(1) :194–244. doi : 10.1353/lan.2015.0005.

T. Gernhard. The conditioned reconstructed process. *Journal of Theoretical Biology*, 253(4) :769–778, Aug. 2008. ISSN 00225193. doi : 10.1016/j.jtbi.2008.04.005. URL `https://linkinghub.elsevier.com/retrieve/pii/S0022519308001811`.

R. D. Gray and Q. D. Atkinson. Language-tree divergence times support the anatolian theory of indo-european origin. 426(6965) :435–439. doi : 10.1038/nature02029.

R. D. Gray, A. J. Drummond, and S. J. Greenhill. Language phylogenies reveal expansion pulses and pauses in pacific settlement. 323(5913) :479–483. doi : 10.1126/science.1166858.

P. Heggarty, C. Anderson, M. Scarborough, B. King, R. Bouckaert, L. Jocz, M. J. Kümmel, T. Jügel, B. Irslinger, R. Pooth, H. Liljegren, R. F. Strand, G. Haig, M. Macák, R. I. Kim, E. Anonby, T. Pronk, O. Belyaev, T. K. Dewey-Findell, M. Boutilier, C. Freiberg, R. Tegethoff, M. Serangeli, N. Liosis, K. Stroński, K. Schulte, G. K. Gupta, W. Haak, J. Krause, Q. D. Atkinson, S. J. Greenhill, D. Kühnert, and R. D. Gray. Language trees with sampled ancestors support a hybrid model for the origin of indo-european languages. 381(6656) :eabg0818. doi : 10.1126/science.abg0818.

P. Jordan and S. O'Neill. Untangling cultural inheritance. 365 (1559) :3875–3888. doi : 10.1098/rstb.2010.0092.

P. Jordan and S. Shennan. Cultural transmission, language, and basketry traditions amongst the california indians. 22 (1) :42–74. doi : 10.1016/S0278-4165(03)00004-7.

R. E. Kass and A. E. Raftery. Bayes factors. 90(430) :773–795. doi : 10.1080/01621459.1995.10476572.

V. Kolipakam, F. M. Jordan, M. Dunn, S. J. Greenhill, R. Bouckaert, R. D. Gray, and A. Verkerk. A bayesian phylogenetic study of the dravidian language family. 5(3) :171504. doi : 10.1098/rsos.171504.

S. Le Bomin, G. Lecointre, and E. Heyer. The evolution of musical diversity. 11(3) :e0151570. doi : 10.1371/journal.pone.0151570.

D. Learmouth, R. H. Layton, and J. J. Tehrani. The evolution of cultural diversity in pama-nyungan australia. 11(945). doi : 10.1057/s41599-024-03386-7.

R. Mace and C. J. Holden. A phylogenetic approach to cultural evolution. 20(3) :116–121. doi : 10.1016/j.tree.2004.12.002.

L. J. Matthews, J. J. Tehrani, F. M. Jordan, M. Collard, and C. L. Nunn. Testing for divergent transmission histories among cultural characters. 6(4) :e14810. doi : 10.1371/journal.pone.0014810.

M. J. O'Brien, M. T. Boulanger, B. Buchanan, R. A. Bentley, R. L. Lyman, C. P. Lipo, M. E. Madsen, and M. I. Eren. Design space and cultural transmission. 23(2) :692–740. doi : 10.1007/s10816-015-9258-7.

S. Passmore, A. L. C. Wood, C. Barbieri, D. Shilton, H. Daikoku, Q. D. Atkinson, and P. E. Savage. Global musical diversity is largely independent of linguistic and genetic histories. 15(3964) :1–12. doi : 10.1038/s41467-024-48113-7.

L. Sagart, G. Jacques, Y. Lai, R. J. Ryder, V. Thouzeau, S. J. Greenhill, and J.-M. List. Dated language phylogenies shed light on the ancestry of sino-tibetan. 116(21) :10317–10322. doi : 10.1073/pnas.1817972116.

C. H. Saslis-Lagoudakis, J. A. Hawkins, S. J. Greenhill, C. A. Pendry, M. F. Watson, W. Tuladhar-Douglas, S. R. Baral, and V. Savolainen. The evolution of traditional knowledge. 281(1780) :20132768. doi : 10.1098/rspb.2013.2768.

M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1) :vey016, June 2018. ISSN 2057-1577. doi : 10.1093/ve/vey016. URL `https://doi.org/10.1093/ve/vey016`. _eprint : https ://academic.oup.com/ve/article-pdf/4/1/vey016/25028516/vey016.pdf.

Y. Tao, Y. Wei, J. Ge, Y. Pan, W. Wang, Q. Bi, P. Sheng, C. Fu, W. Pan, L. Jin, H.-X. Zheng, and M. Zhang. Phylogenetic evidence reveals early kra-dai divergence and dispersal in the late holocene. 14(6924) :1–10. doi : 10.1038/s41467-023-42761-x.

J. J. Tehrani, M. Collard, and S. J. Shennan. The cophylogeny of populations and cultures. 365(1559) :3865–3874. doi : 10.1098/rstb.2010.0020.

Contribution à l'automatisation de l'analyse de mise en page des manuscrits médiévaux

Mise en place d'un test statistique pour décider si une page est vierge

Pierre Lebec[1], Alain Celisse[2], Stéphane Lamassé[1] & Hélène Noizet[1]

[1] *Laboratoire de Médiévistique Occidentale de Paris (UMR 8589), France, pierre.lebec@univ-paris1.fr, stephane.lamasse@univ-paris1.fr et helene.noizet@univ-paris1.fr*
[2] *Statistique, Analyse et Modélisation Multidisciplinaire (UR 4543), France, alain.celisse@univ-paris1.fr*

Résumé. Dans le champ de l'histoire médiévale, une discipline s'intéresse à l'étude matérielle et formelle des manuscrits médiévaux : la codicologie. Jusqu'à présent, les codicologues réalisaient l'intégralité des mesures nécessaires manuellement sur les manuscrits. S'il semble possible d'automatiser tout ou partie de ces relevés, de tels outils restent à développer. Afin de résoudre le problème de l'identification de lignes de textes dans des pages qui n'en contiennent pas, nous avons développé un test statistique permettant de décider si une page est vierge ou non. Ce test s'applique aux pixels d'une image de page de manuscrit et repose sur les modèles de mélange gaussien (GMM). L'une des forces de cette approche est qu'elle parvient à s'adapter aux variations de niveaux de gris entre différentes pages, ce qu permet d'améliorer les résultats par rapports aux approches reposant sur un seuil fixe uniforme sur l'ensemble des numérisations. À l'issue d'une série d'expérimentations préliminaires sur de vraies numérisations (73 pages de manuscrits) libre de toute supervision, nous avons pu observer de très bonnes performances avec des taux de faux positifs et faux négatifs assez faibles.

Mots-clés. Statistique appliquée, test statistique, modèles de mélange gaussien, analyse de mise en page numérique, codicologie, histoire médiévale.

Abstract. In the field of medieval history, codicology is the material and formal study of medieval manuscripts. So far, codicologists have manually gathered all necessary data from manuscripts. Although automating these measurements seems feasible, appropriate tools have yet to be developed. To address the issue of detecting text lines on pages without any text, we developed a statistical test. This test is applied to the pixels of a manuscript page image and is based on Gaussian Mixture Models (GMM). One of the strengths of this approach is its ability to adapt to the grayscale variation between different pages. This adaptability improves the results compared to approaches that use a single fixed threshold for all digitized pages. After a series of preliminary experiments on real digitized pages (73 manuscript pages) conducted without any supervision, we observed excellent performance with low false positive and false negative rates.

Keywords. Applied statistics, statistical testing, Gaussian Mixture Models, digital layout analysis, codicology, medieval history.

1 Introduction

1.1 La codicologie

La codicologie est la discipline historique qui étudie la confection des livres anciens : manuscrits et incunables [1]. D'abord discipline auxiliaire de la science historique, elle s'autonomisa dans les années 1980 sont l'impulsion des travaux du groupe QUANTICOD (1985 et Ornato 1997). Ses membres proposèrent d'introduire des méthodes quantitatives dans ce champ disciplinaire qui ne connaissait alors que des méthodes qualitatives (Muzerelle 1991). Ils revendiquèrent que l'introduction de telles méthodes, largement fondées sur l'usage de statistiques, leur permettraient d'envisager une étude sociale, économique et culturelle de la production livresque au Moyen Âge. Les travaux qu'ils menèrent jusqu'à la fin de leurs carrières dans les années 2000 montrèrent l'efficacité de l'emploi de méthodes statistiques pour écrire cette histoire sociale, économique et culturelle du livre médiéval.

1.2 L'automatisation des relevés et mesures codicologiques

Si le calcul statistique a dès l'origine bénéficié des possibilités offertes par les ordinateurs personnels, une tâche est demeurée purement manuelle : les relevés codicologiques. Les codicologues continuent aujourd'hui de se rendre en centres d'archives ou en bibliothèques afin de consulter des manuscrits et de réaliser eux-mêmes les relevés codicologiques nécessaires à leurs recherches.

Or, ces dix dernières années, l'évolution conjointe des techniques et politiques de conservation des documents historiques d'un côté, et des algorithmes notamment basés sur l'intelligence artificielle (IA) de l'autre, permettent d'envisager la possibilité d'automatiser les relevés et mesures codicologiques. En effet, les institutions de conservations – bibliothèques, centres d'archives, musées – ont rendu et continuent de rendre disponibles, à travers leurs sites en ligne, de nombreux documents numérisés. C'est ainsi plusieurs dizaines de milliers de manuscrits médiévaux numérisés qui y sont disponibles. Concrètement, ce sont des photographies de ces documents qui sont mises à disposition des publics en ligne.

1.3 Une base pour l'analyse de mise page : les solutions de reconnaissance d'écritures

Ces documents disponibles en grandes quantités en ligne au format numériques ne sont pas demeurés inexploités par les historiens. Dans le champ de l'histoire médiévale, depuis la seconde moitié des années 2010, de nombreux projets ont été menés pour d'exploiter ces gisements documentaires numériques afin d'en automatiser la transcription (Daille B. et *al.* 2019, Kiessling 2020, Boillet et *al.* 2022, Clérice et *al.* 2024) [2]. Ces projets relèvent de la

1. Les incunables correspondent aux livres imprimés avant le 1er janvier 1501.

2. La transcription est un travail important dans les recherches en histoire médiévale. Elle consiste en la réécriture de textes manuscrits copiés avec des systèmes d'écriture anciens dans un système d'écriture

reconnaissance optique de caractère (OCR) et plus précisément de la reconnaissance de texte manuscrit (HTR). Afin d'accomplir cette automatisation, ces projets se sont appuyés sur des solutions algorithmiques basées sur l'IA. Dans l'objectif d'optimiser l'efficacité des transcriptions, ils enrichirent leurs outils par l'utilisation d'un ou plusieurs algorithmes préalables réalisant des tâche relevant de l'analyse de mise en page numérique (DLA). Il s'agissait pour eux de détecter les lignes de texte dans une images numérique de page de manuscrit. Reconnaître préalablement la présence de lignes d'écriture et leur position dans l'image améliore sensiblement la précision de la transcription du texte présent sur l'image (Boillet et *al.* 2022).

1.4 Les limites des solutions actuelles pour automatiser les relevés et mesures codicologiques

Si nous disposons déjà de plusieurs outils, comme Kraken ou doc-UFCN, basés sur l'IA, permettant de situer les lignes d'écriture dans une image de page de manuscrit, la précision encore trop faible de ces derniers inhibe la possibilité de les employer seuls pour mener une étude de la mise en page des manuscrits médiévaux. Un problème rencontré couramment est la détection de « lignes fantômes ». Il s'agit de lignes qui ne sont pas réellement sur l'image de page présentée à l'algorithme mais qui sont en réalité au dos de cette page. Lorsque le support d'écriture est suffisamment fin et clair – c'est généralement le cas avec le papier –, il est possible de distinguer par transparence les lignes au dos d'une page. Confrontés à ce phénomène, les algorithmes existants identifient ces « lignes fantômes », perceptibles par transparence, comme des lignes bel et bien présentes sur la page qu'ils analysent.

Nous souhaitions donc développer un moyen de déterminer si une page est vierge ou non, à l'aide d'un algorithme qui ne soit pas dédié à la détection de lignes d'écriture. L'utilisation d'un tel algorithme préalablement à celle d'un algorithme situant les lignes d'écriture pourrait nous permettre d'accroître notre confiance dans les résultats fournis par ce dernier et ainsi améliorer la qualité des données codicologiques exploitées. Pour réaliser cette tâche, nous avons choisi d'élaborer un algorithme dont la décision est fondée sur un test statistique portant sur la distribution du niveau de gris des pixels d'une page de manuscrit numérisée.

2 Type de page et distribution du niveau de gris des pixels

2.1 Étude de la densité du niveau de gris

Si nous avons envisagé de fonder notre algorithme décidant si une page est vierge ou non sur l'étude du niveau de gris des pixels d'une image numérisée de celle-ci, c'est parce que, face à la documentation, nous sommes partis du constat qu'une page vierge devait présenter sensiblement moins de pixels sombres qu'une page sur laquelle quelque chose aurait été copié.

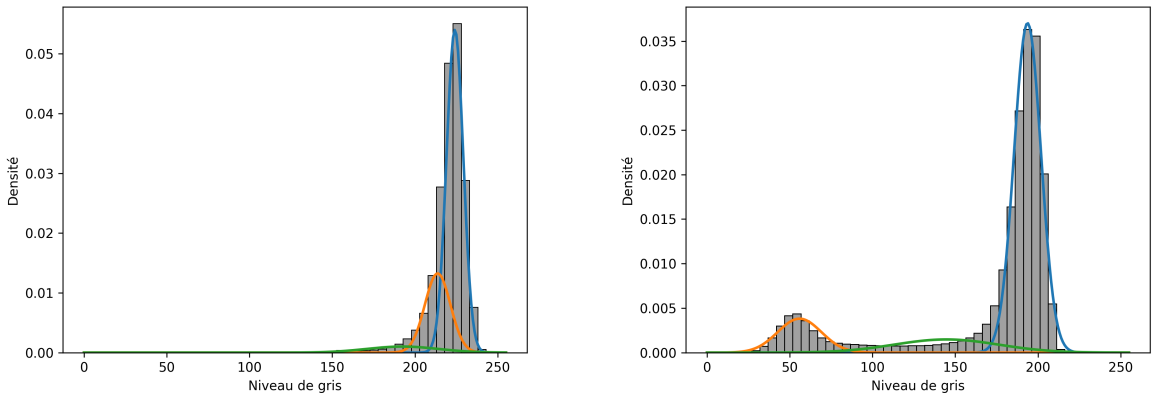
moderne. Cette opération permet de diffuser des sources historiques et les documents sur lesquels s'appuient les argumentaires développés par les historiens.

Afin de vérifier cette intuition, nous avons d'abord observé des distributions du niveau de gris des pixels pour des pages vierges et des pages copiées. La figure 1a présente un exemple d'histogramme pour une page vierge et la figure 1b pour une page copiée. Nous constatons sur cet exemple que ce qui distingue la page vierge de la page copiée c'est l'existence d'un pic autour de la valeur de niveau de gris de 50 pour la page copiée, pic absent dans le cas de la distribution pour la page vierge. Le cas présenté ici est un cas simple, en fonction de la couleur de l'encre et de la densité de l'écriture sur l'image ce pic peut être plus ou moins marqué ou plus ou moins éloignée du pic principal autour de la valeur de 200.

Face à ces distributions, nous avons développé l'intuition qu'elles témoignaient en réalité de la présence de plusieurs sous-populations de pixels qu'il était possible d'assimiler à des objets d'intérêt spécifiques présent sur l'image de la page. Nous pensions que le pic autour de 50 pouvait être associé aux pixels correspondant à une zone copiée de la page ; zone sombre car le support d'écriture est recouvert d'encre (noire ou brune). Et nous pensions que le pic autour de 200 pouvait être associé aux pixels correspondant à la zone demeurée vierge de la page ; zone claire avec une variabilité selon la nature du support : parchemin ou papier.

2.2 Modéliser les distributions à l'aide de mélanges gaussiens

Comme nous envisagions les distributions observées comme la conséquence de la présence, en des proportions variables, de sous-populations de pixels cohérentes associées à des objets d'intérêt, nous avons choisi d'utiliser des modèles de mélanges gaussiens (GMM) pour étudier ces distributions. En premier lieu, s'est posée la question du nombre de composantes à utiliser pour modéliser ces distributions. La figure 1 présente le résultat des modélisations de distribution à l'aide de GMM pour une page vierge et une page copiée.



(a) Histogramme de la distribution observée et GMM la modélisant pour le f. 53v du latin 10266 (page vierge)

(b) Histogramme de la distribution observée et GMM la modélisant pour le f. 5r du latin 16663 (page copiée)

FIGURE 1 – Modélisation de la distribution des pixels pour des images de pages vierge et copiée

3 Décider si une page est vierge ou non : construction d'un test statistique

Nous souhaitions exploiter les GMM afin de mettre en place un test statistique. Pour cela, nous avons utilisé les GMM pour définir une loi sous H_0, avec H_0 : « la page est vierge ». Nous avons défini cette loi sous H_0 comme étant la fonction de répartition F_0 des deux composantes (parmi les trois) correspondant aux sous-populations de pixels les plus clairs. L'objectif du test statistique est de décider si la fonction de répartition du modèle de la distribution observée $\hat{F}$ est significativement distincte ou non de F_0. Notre intuition, fondée sur nos observations, étant que dans le cas d'une page vierge $\hat{F}$ et F_0 sont proches et que dans le cas d'une page copiée elle sont distantes.

3.1 Mesurer la distance entre $\hat{F}$ et F_0 avec les normes L^1 et L^2

La logique du test statistique que nous avons imaginé repose donc sur une mesure de distance entre deux fonctions de répartitions : $\hat{F}$ et F_0. Pour cela nous avons envisagé d'utiliser les normes L^1 et L^2.

3.2 Décision : calcul de la p-valeur

Afin de pouvoir décider si la distance entre $\hat{F}$ et F_0 était suffisamment importante pour décider que la page est copiée (rejet de H_0) ou non, nous avons d'abord défini un seuil de référence S :

$$S = \|\hat{F} - F_0\| \tag{1}$$

Puis, nous avons réitéré notre échantillonnage de pixels. Nous avons choisi de réitérer l'opération à $B = 100$ reprises. Lors de ces réitérations, nous nous sommes intéressés à la variation que connaissait F_0, autrement dit à la variation de la définition de la loi sous H_0. Nous avons défini pour chaque réitération $b \in [1, B]$ un seuil S_b :

$$S_b = \|F_0 - F_0^{(b)}\| \tag{2}$$

avec $F_0^{(b)}$ la fonction de répartition sous la loi H_0 calculée à la réitération b. À partir de ces données, nous avons mis en place le calcul de la p-valeur comme la fréquence des réitérations lors desquelles nous obtenons $S < S_b$:

$$p\text{-valeur} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{(S<S_b)} \tag{3}$$

Suite à nos expérimentations, nous avons conclu que nous rejetions H_0 si p-valeur < 0.05. C'est ce seuil qui nous a donné la meilleure précision pour le test statistique.

4 Évaluation de la précision du test

4.1 Distinction de deux niveaux de vérité-terrain : « strict » et « étendu »

Afin d'évaluer la précision du test statistique, nous avons défini deux niveaux de vérité-terrain. La définition de ces deux niveaux est nécessaire pour interpréter finement nos résultats. Nous avons distingué :

— un niveau « strict », pour lequel seules les pages complètement dépourvues de toute structure tracée à l'encre sur la page sont étiquetées comme vierges ;
— un niveau « étendu », pour lequel toutes les pages qui auraient été considérées comme vierges par un codicologue sont étiquetées comme vierges [3].

Le niveau « strict » correspond donc à une réalité physique : y a-t-il de l'encre sur la page ? Le niveau « étendu » correspond quant à lui à la tâche normalement assurée par un codicologue que nous souhaitons automatiser à l'aide de notre test statistique.

4.2 Comparaison de la précision du test avec les normes L^1 et L^2

Pour mener cette évaluation, nous avons utilisé un échantillon de 73 pages de manuscrits [4] que nous avons étiqueté manuellement. En utilisant la p-valeur calculée pour chacune des pages, le test réalise une prédiction. S'il y a rejet de H_0 au seuil de 5%, alors la page est prédite comme « vierge ». Si au contraire H_0 ne peut être rejetée au seuil de 5%, alors la page est prédite comme « copiée ».

Niveau de vérité-terrain	Taux d'erreur (norme L^1)	Taux d'erreur (norme L^2)
« stricte »	0.08	0.10
« étendue »	**0.04**	0.05

TABLE 1 – Taux d'erreur en fonction de la norme et du niveau de vérité-terrain

La table 1 présente l'évaluation de la précision du test statistique que nous avons mis en place. Nous constatons que l'utilisation de la norme L^1 est légèrement plus efficace que la norme L^2. Nous constatons également que les prédictions de notre test sont plus précises avec le niveau « étendu » de vérité-terrain.

Ainsi, le test nous permet d'approcher efficacement la tâche d'un codicologue ayant à étiqueter des pages comme « vierges » ou « copiées ». Le test se révèle même plus efficace dans cette tâche que dans la discrimination de la réalité physique de la page (présence ou non d'encre sur le support).

3. Dans le cadre d'une étude codicologique portant sur des manuscrits médiévaux, sont ignorés les signes (rappel de cote, estampille, *etc.*) liés à l'archivage moderne, signes généralement apposés sur des pages vierges en début de manuscrit.

4. Il s'agit des manuscrits : latin 5543, latin 7187, latin 10266, latin 14065, latin 16200, latin 16204, latin 16208, latin 16653, latin 16657 et latin 16663 de la Bibliothèque nationale de France (Paris).

5 Conclusion

Le test statistique que nous avons développé afin de décider si une page est vierge ou non à partir de son image numérisée révèle une efficacité qui n'avait, jusqu'alors pas été atteinte pour cette tâche précise. Le fait qu'il ne dépende que des valeurs de niveau de gris des pixels de chaque image de page et non d'une valeur définie à l'échelle du corpus de pages tout entier permet de s'adapter à l'importante variabilité de la documentation médiévale [5]. De surcroît, il est non-supervisé, ce qui évite un coût d'annotation d'un corpus, ce qui le rend plus simple et rapide d'utilisation qu'une solution basée sur un apprentissage supervisé.

Biliographie

Boillet, M., Kermorvant, C. et Paquet, T. (2022), Robust Line Detection in Historical Documents: Learning and Evaluation Methods, *Iternational Journal on Document Analysis and Recognition (IJDAR)*, 25, pp. 95-114.

Clérice T. et *al.* (2024), CATMuS Medieval: A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond, *2024 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 174-194.

Daille, B. et *al.* (2019), Transcription automatique et segmentation thématique de livres d'heures manuscrits, *Revue TAL : traitement automatiques des langues*, 60 (3), pp. 13-36.

Kiessling B. (2020), A Modular Region and Text Line Layout Analysis System, *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp.313-318.

Muzerelle, D. (1991), Evolution et tendances actuelles de la recherche codicologique, *Historia Instituciones Documentos*, 18, pp. 347-374.

Ornato, E. (1997), *La face cachée du livre médiéval. L'histoire du livre vue par Ezio Ornato, ses amis et ses collègues*, La Viella, Rome.

Quanticod (1985), Un programme d'étude quantitative du livre médiéval, *Gaztte du livre médiéval*, 6, pp. 7-13.

5. Par exemple, les encres brunes utilisées dans les manuscrits carolingiens (VIII^{e}-IX^{e} siècle) ou dans certains manuscrits du XV^{e} siècle ne donnent pas les mêmes distributions de niveau de gris des pixels que les encres noires des manuscrits gothiques (XII^{e}-XV^{e} siècle). Il en va de même pour le type d'écriture : caroline, *textualis* gothique, bâtarde gothique, cursive gothique, *etc.*

Statistique appliquée à la biologie-santé

Apprentissage de l'espace acoustique des pleurs de bébé

Guillem Bonafos [a,b,1], Rémi Emonet [a,d,2], Lény Lego [b,4], Nicolas Mathevon [b,d,e,6], Hugues Patural [c,5] & Jérémy Rouch [b,3]

[a] *Université Jean Monnet Saint-Etienne, CNRS, Institut d'Optique Graduate School, Inria, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT- ETIENNE, France*
[b] *ENES Bioacoustics Research Lab, CRNL, University of Saint-Etienne, CNRS, Inserm, Saint-Etienne, France*
[c] *Neonatal and Pediatric Intensive Care Unit, University of Saint-Etienne, SAINBIOSE Laboratory, Inserm, F-42023, Saint-Etienne, France*
[d] *Institut universitaire de France*
[e] *Ecole Pratique des Hautes Etudes, CHArt Lab, PSL University, Paris, France*

prenom.nom@univ-st-etienne.fr

Résumé. Nous traitons dans cette présentation du transfert d'apprentissage de modèles de langage pré-entrainé de manière auto-supervisé pour l'analyse de pleurs de bébés. Nous présentons des résultats empiriques démontrant que les représentations latentes de ces modèles permettent de classifier efficacement les pleurs de bébé. Pour cela, nous utilisons les représentations de différents modèles sur des tâches de classification supervisées de pleurs. Nous analysons plus en détail ces résultats, ce qui nous permet de déduire l'information acoustique contenue dans ces représentations.

Mots-clés. apprentissage, représentation, transfert d'apprentissage, pleur, bébé

Abstract. In this presentation, we explore the use of pre-trained language models in a self-supervised manner for analyzing infant cries. We present empirical evidence showing that the latent representations produced by these models can effectively support the classification of infant cries. To achieve this, we leverage representations from various models on supervised cry classification tasks. We then conduct a detailed analysis of the results, which enables us to infer the acoustic information embedded within these representations.

Keywords. learning, representation, transfer learning, cry, baby

1 Description de la communication

1.1 Contexte global

Les pleurs des bébés, en plus d'être un signal de communication, sont également une source importante d'informations : ils encodent une signature individuelle permettant de

reconnaitre l'enfant qui pleure et son âge (Lockhart-Bouron *et al.*, 2023), le niveau de détresse et de douleur (Koutseff *et al.*, 2018). Ils ont également un potentiel diagnostic (Kheddache et Tadj, 2019), mais les preuves établissant un lien entre pleur et troubles potentiels ne sont pas encore suffisamment solides pour pouvoir conclure (Lawford *et al.*, 2022).

Pour dépasser cette limite, une base de données massives et de qualité, visant à enregistrer les pleurs de mille bébés, est en cours d'enregistrement. Ces enfants, enregistrés de leur naissance jusqu'à leur sortie de la maternité, sont revus deux ans plus tard afin de leur faire passer le test Bailey, une batterie permettant d'évaluer le fonctionnement développemental de l'enfant. L'objectif du projet de recherche est d'estimer la capacité prédictive des pleurs sur des parties du test.

Cette présentation s'inscrit dans ce projet plus global et présente les premières avancées du travail statistique qu'il implique, à savoir l'apprentissage d'une représentation du pleur.

1.2 Motivations

Classiquement, l'analyse des pleurs passe par le calcul de descripteurs acoustiques tels que leur fréquence fondamentale, leurs formants, leur rugosité, etc. L'apprentissage de représentation, rendue possible par la construction de bases de données massives, permet de se passer du calcul de tels descripteurs et d'avoir une représentation construite sur les données.

Cela ouvre notamment la porte à des modèles génératifs tels que les autoencodeurs variationnels (Kingma et Welling, 2014), à partir desquels il est possible de faire un lien avec la théorie source-filtre (Sadok *et al.*, 2023). On peut donc potentiellement extraire plus d'information des pleurs, et ainsi avoir une meilleure capacité de prédiction, tout en conservant un potentiel explicatif. Les résultats récents dans le domaine du speech motivent également à apprendre une représentation du pleur. En effet, les modèles basés sur des modules d'attention (Vaswani *et al.*, 2017) comme wav2vec (Schneider *et al.*, 2019; Baevski *et al.*, 2020), ont permis de grandement améliorer les résultats et sont devenus l'état de l'art. Leur stratégie repose sur l'apprentissage de représentations solides du signal vocal par des méthodes auto-supervisées.

Si l'apprentissage de représentation a démontré son intérêt et est aujourd'hui l'état de l'art dans de nombreux domaines, son succès repose en grande partie sur la quantité de données à disposition Bengio *et al.* (2013). C'est le cas des modèles de langage naturel, qui s'appuient sur des quantités de données d'une toute autre échelle que celles à disposition pour l'analyse des pleurs de bébés. Le transfert d'apprentissage peut alors être une solution pour dépasser cette limite.

1.3 La question traitée

Nous présentons des résultats empiriques démontrant l'intérêt du transfert d'apprentissage de modèles de langage naturel pour l'analyse du pleur de bébés. Nous utilisons également les résultats obtenus sur les différentes tâches afin d'inférer l'information contenue dans les représentations latentes des modèles.

Nous listons l'ensemble des datasets disponibles dans le domaine, que nous avons centralisé. Chacun de ces datasets contient des pleurs de bébés, avec des étiquettes correspondantes, chaque ensemble d'étiquette permettant de définir une tâche de classification. Au total, les huit datasets utilisés permettent de définir 23 tâches de classification. Nous utilisons les représentations latentes de cinq modèles de langage pour vectoriser chaque pleur de bébé. Nous comparons ensuite les résultats pour l'ensemble des tâches de classification et représentations latentes. Nous utilisons le même modèle de classification pour chacune, à savoir une forêt aléatoire de 150 arbres. Nous entrainons ainsi $23 \times 5 = 115$ forêts aléatoires.

L'analyse des résultats obtenus sur chaque tâches de classification (Figure 1 nous montre que ces modèles pré-entrainés sur des ondes sonores de données langagières permettent bel et bien d'extraire une information des pleurs utile pour des tâches de prédiction. Plus particulièrement, nous montrons que nous sommes capables de prédire l'identité de l'enfant qui pleure ou si le pleur est induit par une douleur ou non. Nous apportons ainsi un premier résultat sur l'intérêt de l'approche pour le domaine. L'absence de résultats obtenus pour certaines tâches de classification, comme par exemple l'incapacité de prédire le sexe de l'enfant pleureur, est cohérente avec la littérature, étant donné qu'il n'y a pas à cet âge là de différence dans la structure acoustique des pleurs.

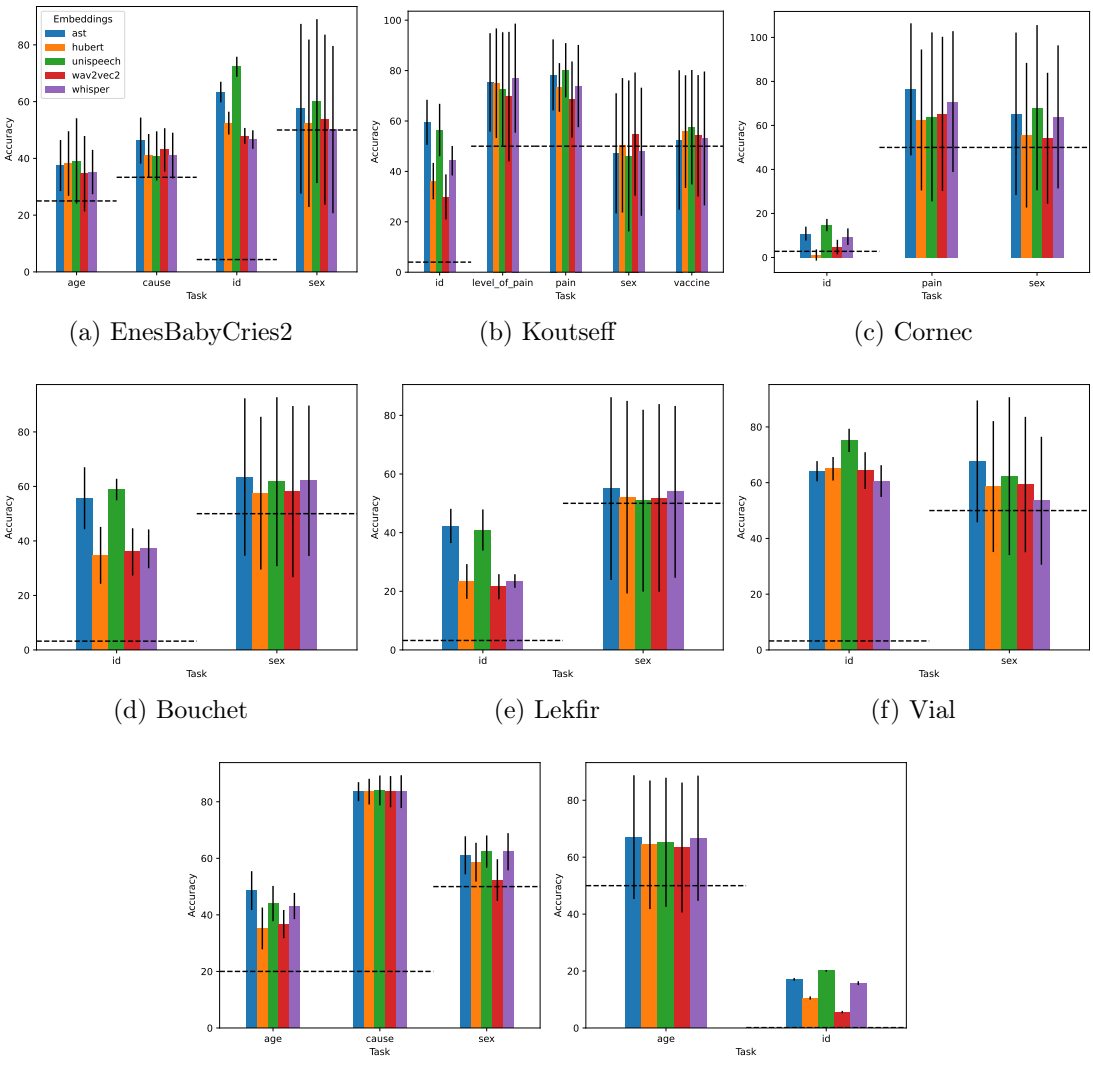
Nous faisons ensuite le lien entre la représentation apprise et des descripteurs acoustiques plus classiques. Premièrement, la capacité de ces modèles pré-entrainés sur du langage à extraire une information des pleurs utile pour des tâches de classification est informative quant à l'information contenue dans leurs représentations latentes. Deuxièmement, l'analyse de la distribution des pleurs dans cet espace (Figure 2), en lien avec les valeurs prises par ces descripteurs, permet d'en apprendre plus sur l'information encodée par la représentation.

1.4 Intérêt et nouveauté

Nous présenterons une application nouvelle et encore inexplorée de l'apprentissage, à savoir le transfert d'apprentissage par l'utilisation de représentations latentes de modèles pré-entrainés de manière auto-supervisé pour la classification de pleurs de bébé, et la question de l'interprétabtilité des représentations latentes. Nous compilons pour cela des datasets divers et nombreux, et nous appliquons la même méthodologie pour produire des résultats empiriques robustes et réplicables.

Cette application sur de la classification de pleurs d'enfant permet de mettre en lumière les paramètres acoustiques encodés dans les représentations latentes de ces modèles. Dans le cas de la douleur par exemple, l'association du pleur de douleur à des phénomènes acoustiques non-linéaires est maintenant bien établi dans la littérature. C'est la rugosité d'un pleur, et pas une autre caractéristique acoustique, qui encode la douleur. Aussi, si une représentation permet de prédire qu'un pleur est un pleur de douleur, cela veut dire que cette représentation encode la rugosité de la vocalisation. L'application de ces modèles à ce domaine permet ainsi d'en apprendre plus sur les représentations latentes de ces modèles, souvent perçus comme des boites noires.

Concernant le versant applicatif, ce travail doit permettre de répondre à des questions



(a) EnesBabyCries2 (b) Koutseff (c) Cornec

(d) Bouchet (e) Lekfir (f) Vial

(g) Donate A Cry (h) Cry Celeb

FIGURE 1 – Exactitude (*accuracy*) pour chaque base de données et chaque embedding (barres d'erreur = écart type ; ligne horizontale = niveau du hasard).

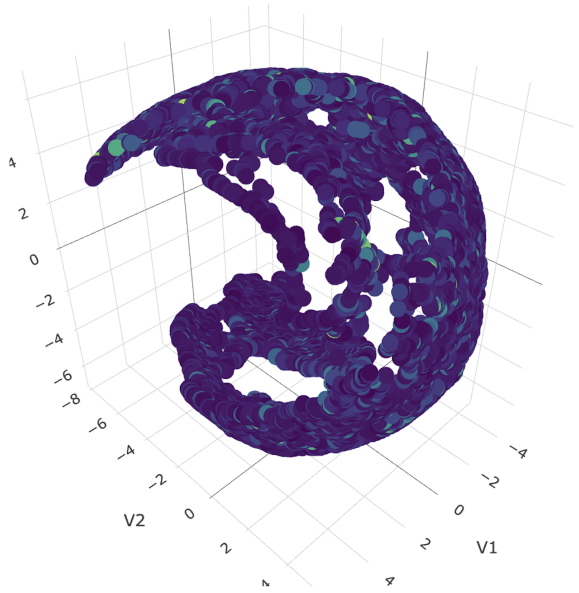


FIGURE 2 – Distribution des pleurs dans l'espace latent appris. Réduction de l'espace latent à 3 dimensions par Uniform Manifold Approximation and Projection (UMAP)

spécifiques en biologie et santé : est-il possible d'utiliser le pleur comme outil diagnostic ? Le pleur encode-t-il une information prédictive du développement à venir de l'enfant ? Y a-t-il une différence entre l'information perceptible et contenue dans le pleur ?

Bibliographie

Alexei BAEVSKI, Yuhao ZHOU, Abdelrahman MOHAMED et Michael AULI : wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations. *In Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html`.

Yoshua BENGIO, Aaron COURVILLE et Pascal VINCENT : Representation Learning : A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, août 2013. ISSN 1939-3539. Conference Name : IEEE Transactions on Pattern Analysis and Machine Intelligence.

Yasmina KHEDDACHE et Chakib TADJ : Identification of Diseases in Newborns Using Advanced Acoustic Features of Cry Signals. *Biomedical Signal Processing and Control*, 50:35–44, 2019. ISSN 1746-8094. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7672377/`.

Diederik P. KINGMA et Max WELLING : Auto-Encoding Variational Bayes. *In*

arXiv :1312.6114 [cs, stat]. Conference proceedings : papers accepted to the International Conference on Learning Representations (ICLR) 2014, mai 2014. URL http://arxiv.org/abs/1312.6114. arXiv : 1312.6114.

Alexis Koutseff, David Reby, Olivier Martin, Florence Levrero, Hugues Patural et Nicolas Mathevon : The acoustic space of pain : cries as indicators of distress recovering dynamics in pre-verbal infants. *Bioacoustics*, 27(4):313–325, octobre 2018. ISSN 0952-4622, 2165-0586. URL https://www.tandfonline.com/doi/full/10.1080/09524622.2017.1344931.

Harriet L. S. Lawford, Hannah Sazon, Céline Richard, Michael P. Robb et Samudragupta Bora : Acoustic Cry Characteristics of Infants as a Marker of Neurological Dysfunction : A Systematic Review and Meta-Analysis. *Pediatric Neurology*, 129:72–79, avril 2022. ISSN 1873-5150.

Marguerite Lockhart-Bouron, Andrey Anikin, Katarzyna Pisanski, Siloé Corvin, Clément Cornec, Léo Papet, Florence Levréro, Camille Fauchon, Hugues Patural, David Reby et Nicolas Mathevon : Infant cries convey both stable and dynamic information about age and identity. *Communications Psychology*, 1(1):1–15, octobre 2023. ISSN 2731-9121. URL https://www.nature.com/articles/s44271-023-00022-z. Number : 1 Publisher : Nature Publishing Group.

Samir Sadok, Simon Leglaive, Laurent Girin, Xavier Alameda-Pineda et Renaud Séguier : Learning and controlling the source-filter representation of speech with a variational autoencoder. *Speech Communication*, 148:53–65, mars 2023. ISSN 0167-6393. URL https://www.sciencedirect.com/science/article/pii/S0167639323000304.

Steffen Schneider, Alexei Baevski, Ronan Collobert et Michael Auli : wav2vec : Unsupervised Pre-Training for Speech Recognition. *In Interspeech 2019*, pages 3465–3469. ISCA, septembre 2019. URL https://www.isca-archive.org/interspeech_2019/schneider19_interspeech.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser et Illia Polosukhin : Attention is All you Need. *In Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Méthode ADD MALTS pour l'inférence causale avec outcome distributionnel : évaluation d'un programme d'hypertension sur tensiomètres connectés Withings

Agathe Chabassier [1] & Julie Josse [2] & Erwan Scornet [3] & Claire Boyer [4] & Pierre Binon [5]

[1] *Inria Montpellier, PreMedical - Sorbonne Université, LPSM - Withings ; France ; agathe.chabassier@withings.com*
[2] *Inria Montpellier, PreMedical; France ; julie.josse@inria.fr*
[3] *Sorbonne Université, LPSM ; France ; erwan.scornet@polytechnique.edu*
[4] *Sorbonne Université, LPSM ; France ; claire.boyer@sorbonne-universite.fr*
[5] *Withings ; France ; pierre.binon@withings.com*

Résumé. Ce rapport présente une application de la méthode ADD MALTS, une approche novatrice proposée par Katta et al. (2024) pour l'inférence causale de l'effet d'un traitement binaire sur un outcome distributionnel issu de capteurs. L'application considérée concerne l'impact de la participation à un programme d'hypertension de deux mois sur la distribution des valeurs de tensions mesurées le troisième mois, par rapport à des individus contrôles ayant un tensiomètre mais ne faisant pas partie du programme. Plutôt que de résumer les mesures répétées par une statistique scalaire (comme la moyenne), la méthode ADD MALTS exploite la distribution complète des mesures, permettant de capturer la variabilité ainsi que les extrêmes, généralement cliniquement pertinents. Dans un premier temps, une distance optimale entre les covariables confondantes des utilisateurs est calculée sur un ensemble d'entraînement. Ensuite, pour chaque utilisateur de l'ensemble d'estimation, ses K plus proches voisins de traitement opposé selon cette distance sont identifiés, et le barycentre de Wasserstein de leurs outcomes distributionnels sert d'approximation de l'outcome contrefactuel de l'utilisateur. L'effet individuel du traitement (ITE) est alors estimé comme la différence entre les quantiles des distributions observée et contrefactuelle, et l'effet moyen du traitement (ATE) pour chaque quantile est obtenu en moyennant les effets individuels. Cette approche permet d'ajuster simultanément sur des covariables statiques telles que l'âge, le sexe ou l'IMC et sur la distribution initiale de la pression artérielle, tout en détectant les violations de positivité. Les résultats montrent que le programme exerce un effet causal en diminuant la pression artérielle systolique après deux mois par rapport à la simple possession d'un tensiomètre Withings, essentiellement chez les utilisateurs ne prenant pas de médicaments antihypertenseurs.

Mots-clés. Inférence causale ; Matching ; Données distributionnelles ; Distance de Wasserstein ; Hypertension.

Abstract. This report presents an application of the ADD MALTS method, an innovative approach proposed by Katta et al. (2024) for causal inference of the effect of a binary treatment on a distributional outcome derived from sensor data. The application considered focuses on the impact of participation in a two-month hypertension program on the distribution of blood pressure values measured in the third month, compared to control individuals

who own a blood pressure monitor but are not part of the program. Rather than summarizing repeated measurements using a scalar statistic (such as the mean), the ADD MALTS method leverages the entire distribution of measurements, allowing it to capture variability as well as extremes, which are often clinically relevant. First, an optimal distance between the users' confounding covariates is learned on a training set. Then, for each user in the test set, their K nearest neighbors with the opposite treatment assignment are identified based on this distance, and the Wasserstein barycenter of their distributional outcomes is used as an approximation of the user's counterfactual outcome. The Individual Treatment Effect (ITE) is then estimated as the difference between the quantiles of the observed and counterfactual distributions, and the Average Treatment Effect (ATE) for each quantile is obtained by averaging the individual effects. This approach enables simultaneous adjustment for static covariates such as age, sex, or BMI and for the initial distribution of blood pressure, while also detecting positivity violations. The results show that the program exerts a causal effect in lowering systolic blood pressure after two months compared to merely owning a Withings blood pressure monitor, with the effect being primarily observed among users not taking antihypertensive medication.

Keywords. Causal inference ; Matching ; Distributional data ; Wasserstein distance ; Hypertension.

1 Introduction

Les dispositifs portables de santé offrent désormais la possibilité de suivre en continu des mesures physiologiques telles que la pression artérielle. Les approches classiques d'inférence causale s'appuient souvent sur des agrégations scalaires de ces données, perdant ainsi une information cruciale sur la variabilité et les extrêmes, qui sont déterminants pour la gestion du risque clinique.

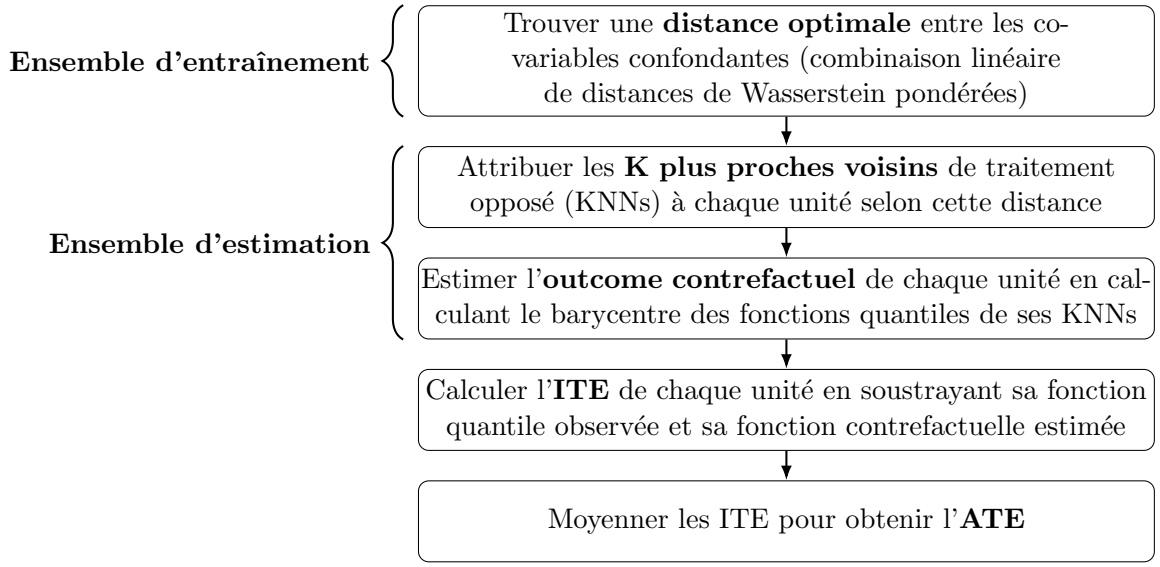
Cette étude est une **application et une extension de la méthode ADD MALTS présentée par Katta et al. (2024)**. En s'appuyant sur la distance de Wasserstein W_2 et le barycentre de Wasserstein, l'approche étudiée représente l'outcome par la fonction quantile de ses valeurs $F_Y^{-1}(q)$ et exploite l'ensemble de la distribution pour estimer les effets du traitement (le programme d'hypertension).

2 Méthodologie

2.1 Vue d'ensemble de la méthode ADD MALTS

La méthode ADD MALTS consiste à matcher chaque individu à des individus lui ressemblant ayant eu le traitement opposé, afin d'inférer son outcome contrefactuel (qu'il aurait eu s'il avait eu le traitement opposé), avec des outcomes et covariables qui sont des distributions de mesures et non de simples scalaires.

Pour identifier les effets causaux, les hypothèses suivantes sont posées : **SUTVA** (les outcomes potentiels, ici des distributions, dépendent uniquement du traitement appliqué à chaque unité), l'**ignorabilité conditionnelle** (les outcomes potentiels sont indépendants du traitement, conditionnellement aux variables confondantes F_{X_i}) et la **positivité** (chaque unité a une probabilité strictement comprise entre 0 et 1 d'être traitée).



2.2 Modélisation des issues distributionnelles

Pour chaque individu i, l'outcome est représenté par une fonction de répartition notée F_{Y_i}, qui correspond à la distribution des mesures de pression artérielle observées pendant le troisième mois après initiation du programme. La fonction quantile associée, $F_{Y_i}^{-1}(q)$, fournit la valeur minimale de tension telle que la probabilité d'observer une mesure inférieure soit au moins q. Les covariables confondantes, regroupées sous la notation F_{X_i}, incluent à la fois des caractéristiques statiques (considérées comme des distributions dégénérées) — telles que l'âge ou le sexe — et des informations de nature distributionnelle, ici la distribution de pression artérielle le mois précédant l'inclusion.

La **distance entre deux distributions**, μ et ν, est mesurée par la distance de Wasserstein d'ordre 2 :

$$W_2(\mu, \nu) = \left(\int_0^1 \left| \mu^{-1}(q) - \nu^{-1}(q) \right|^2 dq \right)^{1/2}.$$

Le **barycentre de Wasserstein**, utilisé pour agréger plusieurs distributions, est défini par la formule de gauche ci-dessous, et égal à celle de droite pour des distributions continues et unidimensionnelles (Lin et al., 2022) :

$$B[F_Y] := \arg\min_{F \in \mathcal{W}_2(I)} E\left[W_2^2\left(F_Y, F\right)\right] \quad \text{et} \quad B[F_Y]^{-1}(q) = E\left[F_Y^{-1}(q)\right].$$

2.3 Première étape : calcul d'une distance pertinente entre les covariables confondantes

La première étape consiste à **trouver des coefficients optimaux dans la distance entre les covariables distributionnelles** des individus i et j, définie comme une combinaison linéaire de distances de Wasserstein :

$$d_M(F_{X_i}, F_{X_j}) = \sum_{l=1}^{d} M_{l,l}\, W_2^2\left(F_{X_{i,l}}, F_{X_{j,l}}\right),$$

où $F_{X_{i,l}}$ représente la $l^{\text{ème}}$ composante (statique ou distributionnelle) du vecteur de covariables de l'individu i et M est une matrice diagonale de pondération.

L'optimisation de M s'effectue sur un **ensemble d'entraînement** S_{tr} en minimisant :

$$M^*(S_{\text{tr}}) \in \arg\min_M \Big(c\|M\|_F + \Delta^{(1)}(M) + \Delta^{(0)}(M)\Big),$$

avec

$$\Delta^{(t)}(M) = \frac{1}{|S_{\text{tr}}^{(t)}|} \sum_{i \in S_{\text{tr}}^{(t)}} W_2^2(\hat{F}_{Y_i,M}, F_{Y_i}),$$

où $\hat{F}_{Y_i,M}$ est l'estimation de la fonction de répartition de l'unité i obtenue en prenant le barycentre des K plus proches voisins (KNN) ayant le même traitement.

2.4 Deuxième étape : calcul des outcomes contrefactuels et estimation

Il est important de noter que le calcul de la distance est fait sur un ensemble d'entraînement S_{tr} différent de l'**ensemble d'estimation** S_{est} sur lequel sont effectués le matching et l'estimation des effets afin éviter des problèmes de "data leakage".

L'**effet moyen conditionnel du traitement (CATE)** est ensuite estimé comme la différence entre la fonction quantile observée pour l'unité i et la fonction quantile contrefactuelle. Formellement, on a, pour un individu ayant reçu le traitement $T_i = 1$, avec une formule similaire pour $T_i = 0$:

$$\hat{\tau}(q \mid F_{X_i}) = F_{Y_i}^{-1}(q) - \hat{B}[F_Y \mid F_X = F_{X_i}, T = 0]^{-1}(q)$$

Les estimands causaux sont définis de la manière suivante. L'effet individuel du traitement (ITE) pour une unité i est mesuré par la différence entre la fonction quantile du potentiel résultat sous traitement et celle sous contrôle, c'est-à-dire $\tau_i(q) = F^{-1}_{Y_i(1)}(q) - F^{-1}_{Y_i(0)}(q)$ pour tout $q \in [0,1]$. Le CATE correspond à l'espérance conditionnelle de l'ITE, soit $\tau(q \mid F_X) = \mathbb{E}\left[\tau_i(q) \mid F_{X_i} = F_X\right]$, tandis que l'effet moyen du traitement (ATE) est obtenu en moyennant les ITE sur l'ensemble de la population, c'est-à-dire $\tau(q) = \mathbb{E}\left[\tau_i(q)\right]$.

3 Application au Programme d'Hypertension de Withings

3.1 Contexte et constitution des cohortes

Le programme d'hypertension de Withings est un **programme comportemental visant à aider les utilisateurs à réguler leur tension par des conseils adaptés**. Les données proviennent de tensiomètres connectés Withings, qui permettent de mesurer la pression artérielle de manière régulière. L'objectif est d'évaluer l'impact de ce programme sur la distribution des valeurs de pression artérielle deux mois après l'initiation du programme, par rapport à des individus contrôles ayant un tensiomètre mais ne faisant pas partie du programme.

Deux cohortes ont été définies. La première, la **cohorte des utilisateurs du programme**, regroupe les patients ayant démarré le programme (la plupart après mars 2023) et disposant d'au moins une mesure de pression artérielle durant le mois précédant le début du programme ainsi que d'une mesure deux mois après. La seconde, la **cohorte témoin**, rassemble les patients n'ayant jamais participé au programme, avec une date d'inclusion artificielle fixée au 1er mars 2023 et respectant les mêmes critères de disponibilité des mesures. Dans les deux cas, les utilisateurs doivent avoir répondu à un questionnaire notamment sur leur prise de médicaments antihypertenseurs dans les 10 jours après leur inclusion. Les deux cohortes présentent des profils similaires quant aux moments de mesure, caractérisés par une distribution bimodale avec des pics d'enregistrement le matin et en soirée. Les nombres d'utilisateurs et de contrôles dans les ensembles d'entraînement et d'estimation sont indiqués dans le tableau 1.

Ensemble	**Programme**	**Contrôles**	**Total**
Entraînement	249	334	583
Estimation	1080	1256	2336
Total	1329	1590	2919

Table 1: Nombre d'utilisateurs et de contrôles dans les ensembles d'entraînement et d'estimation.

3.2 Schéma causal et ajustements par la méthode ADD MALTS

Via le matching effectué, la méthode ADD MALTS permet d'ajuster simultanément sur des covariables statiques (âge, sexe, IMC, date de première mesure de pression artérielle, nombre de mesures initiales comme proxy de la motivation des utilisateurs) et sur la distribution initiale de la pression artérielle un mois avant l'inclusion (début du programme pour les utilisateurs et date artificielle d'inclusion le 1er mars 2023 pour les contrôles). Les variables sont normalisées avant ajustement. **Les covariables d'ajustements ont été choisies après discussion avec des médecins et des spécialistes du programme d'hypertension**

chez Withings. Par ailleurs, plusieurs ensembles d'ajustement ont été testés et donnent des résultats très similaires, ce qui valide la robustesse de l'approche. La Figure 1 illustre les ajustements effectués.

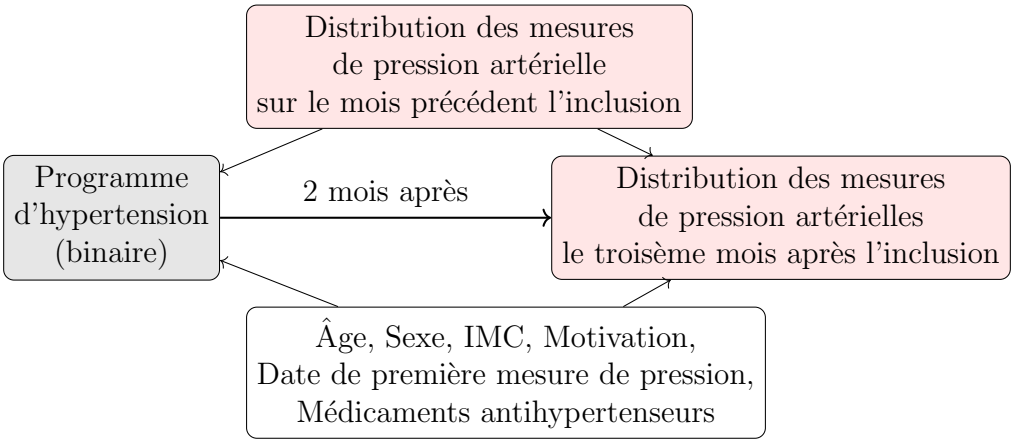


Figure 1: Schéma des ajustements effectués. Le traitement est grisé, les variables distributionnelles apparaissent en rouge et les covariables scalaires en blanc.

Les pondérations obtenues pour la distance de Wasserstein entre les covariables confondantes à l'étape 1 sont présentées dans la Table 2. On note que **la prise de médicaments antihypertenseurs est la variable la plus influente dans le calcul de la distance, suivie par la motivation** (nombre de mesures le mois avant l'inclusion), comme on pouvait s'y attendre car ces deux variables sont fortement prédictives de l'outcome (la pression artérielle après deux mois).

Variable	**Coefficient**
Âge	0.062
Sexe	0.073
IMC initial	0.011
Prise de médicaments antihypertenseurs	**0.60**
Date de la première mesure de pression artérielle	0.17
Motivation (nombre de mesures le mois avant l'inclusion)	**0.27**
Distribution de pressions artérielles initiale	0.043

Table 2: Pondérations obtenues pour la distance de Wasserstein entre les covariables confondantes.

La figure 2 permet de **vérifier que l'ajustement sur les covariables statiques a bien permis d'obtenir deux groupes comparables** en terme d'âge, sexe, IMC, motivation, date de première mesure de pression et prise de médicaments antihypertenseurs.

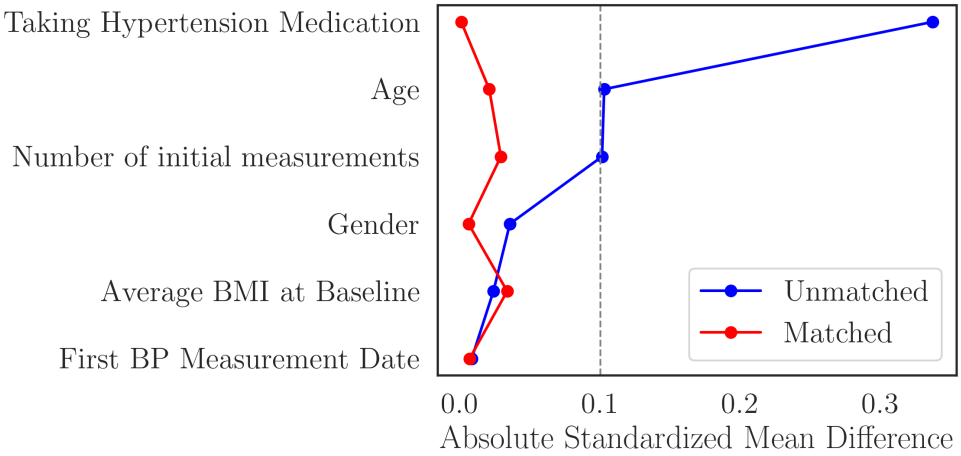


Figure 2: Love plot avec les différences moyennes standardisées entre les utilisateurs du programme et les contrôles pour les variables confondantes statiques avant et après matching.

3.3 Résultat d'estimation

La Figure 3 représente l'effet causal du programme d'hypertension, après ajustement des biais confondants effectué par la méthode ADD MALTS. Cette approche a permis d'**équilibrer avec succès les groupes comparés** non seulement sur des variables statiques (âge, sexe, etc.), comme le prouve le love plot de la Figure 2, mais également sur les distributions initiales de la tension artérielle, comme le montre la superposition des distributions initiales de pression (en pointillés) dans la Figure 3. Cet ajustement rigoureux réduit les biais confondants et isole ainsi **l'impact du programme, qui se traduit par une diminution de la pression artérielle systolique sur deux mois**.

Par ailleurs, il apparaît que **cet effet causal est particulièrement marqué chez les utilisateurs ne prenant pas de médicaments antihypertenseurs**, ce qui était attendu dans la mesure où, en l'absence de traitement pharmacologique, les interventions comportementales jouent un rôle plus déterminant dans la régulation de la pression artérielle (Figure 4).

Les diminutions observées (2 à 3 mmHg) sont comparables, voire prometteuses, au regard d'autres études portant sur l'effet de programmes comportementaux sur la pression artérielle, et ce après seulement deux mois d'intervention (Rodriguez et al., 2021). Notons par ailleurs l'effet observé est mesuré par rapport à la simple possession d'un tensiomètre, laquelle peut à elle seule contribuer à une baisse de la pression artérielle. Cependant, une évaluation sur des périodes plus longues est nécessaire pour confirmer la persistance de ces effets.

4 Conclusion et perspectives

Cette étude confirme l'intérêt d'une approche distributionnelle pour l'inférence causale en santé connectée. En travaillant directement sur les fonctions de répartition et en exploitant

la géométrie du Wasserstein, l'approche ADD MALTS permet de capturer des effets de traitements binaires sur des distributions en ajustant simultanément sur des covariables statiques et distributionnelles.

Ainsi, cette méthode a permis de mettre en évidence l'impact positif du programme d'hypertension de Withings sur la pression artérielle au bout de deux mois, en particulier chez les utilisateurs ne prenant pas de médicaments antihypertenseurs.

D'un point de vue méthodologique, **la méthode ADD MALTS présente plusieurs atouts majeurs**. Elle offre une interprétabilité à la fois individuelle et populationnelle (l'ATE étant défini comme la moyenne des ITEs), garantit la cohérence avec l'ATE classique dans le cas de distributions dégénérées, et fournit un estimateur consitent plus performant que les principales approches existantes en analyse distributionnelle pour l'estimation d'effets de traitement. Par ailleurs, elle permet d'identifier les variables les plus confondantes — c'est-à-dire celles qui prédisent le mieux l'outcome — et gère conjointement des covariables de nature scalaire ou distributionnelle.

Plusieurs **axes d'amélioration** sont envisagés pour renforcer cette méthode. D'une part, le développement d'intervalles de confiance permettrait de mieux quantifier l'incertitude des estimations. D'autre part, son extension à des contextes multidimensionnels pourrait en élargir le champ d'application. L'intégration d'une approche de type IPCW (Inverse Probability of Censoring Weights) serait également pertinente pour corriger le biais de sélection des utilisateurs les plus motivés ayant deux mois de données disponibles. En parallèle, il serait utile de mieux prendre en compte l'échantillonnage irrégulier dans la construction des fonctions quantiles — par exemple, en pondérant davantage les mesures isolées — ainsi que d'améliorer la gestion des données manquantes, actuellement imputées par la moyenne. Enfin, l'adaptation de la méthode à des traitements multicatégoriels ou continus constitue une piste prometteuse. Bien que la méthode ADD MALTS se montre déjà plus performante que les principales approches d'analyse distributionnelle pour l'estimation d'effets de traitement, une comparaison systématique avec d'autres méthodes d'inférence causale sur données distributionnelles — notamment celles visant des estimands scalaires plutôt que fonctionnels, comme ce pourrait être le cas d'une extension de la méthode proposée par Even et Josse (2025) — serait particulièrement éclairante.

Bibliographie

- Katta, S. et al. (2024). *Interpretable Causal Inference for Analyzing Wearable, Sensor, and Distributional Data*. arXiv, http://arxiv.org/abs/2312.10569.
- Lin, Z. et al. (2022). *Causal Inference on Distribution Functions*. arXiv, https://doi.org/10.48550/arXiv.2101.01599.
- Even, M., and Josse, J. (2025). *Rethinking the Win Ratio: A Causal Framework for Hierarchical Outcome Analysis*. arXiv, https://doi.org/10.48550/arXiv.2501.16933.
- Rodriguez, M. A. et al. (2021). *Sustained Benefit of Alternate Behavioral Interventions to Improve Hypertension Control*. Hypertension, 77(6), 1867–1876. https://doi.org/10.1161/HYPERT

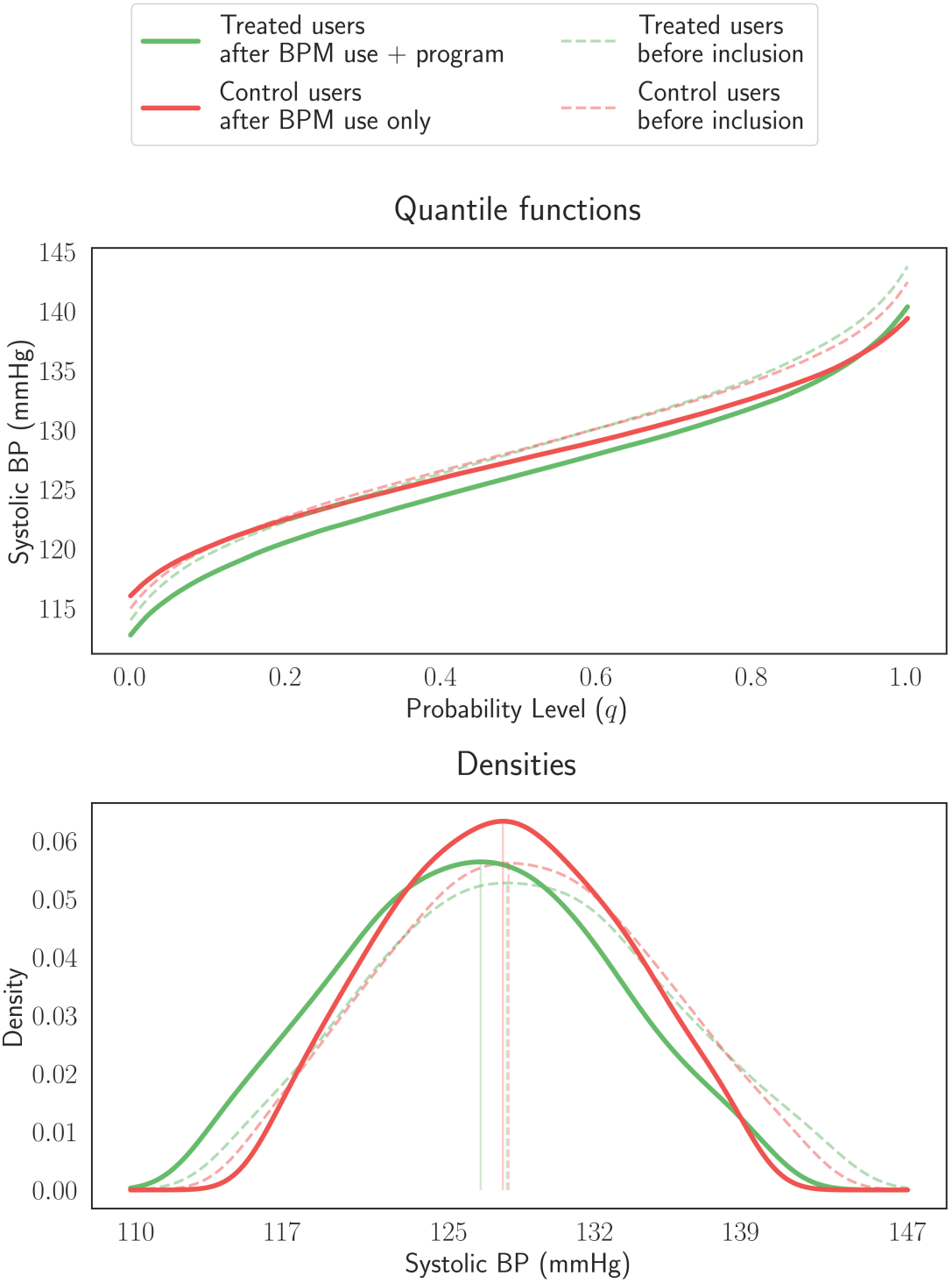


Figure 3: Fonction quantiles moyennes des utilisateurs du programme (en vert) et des contrôles (en rouge) après matching (en haut), et densités de pressions artérielles associées (en bas). NB: BPM: Blood Pressure Monitor.

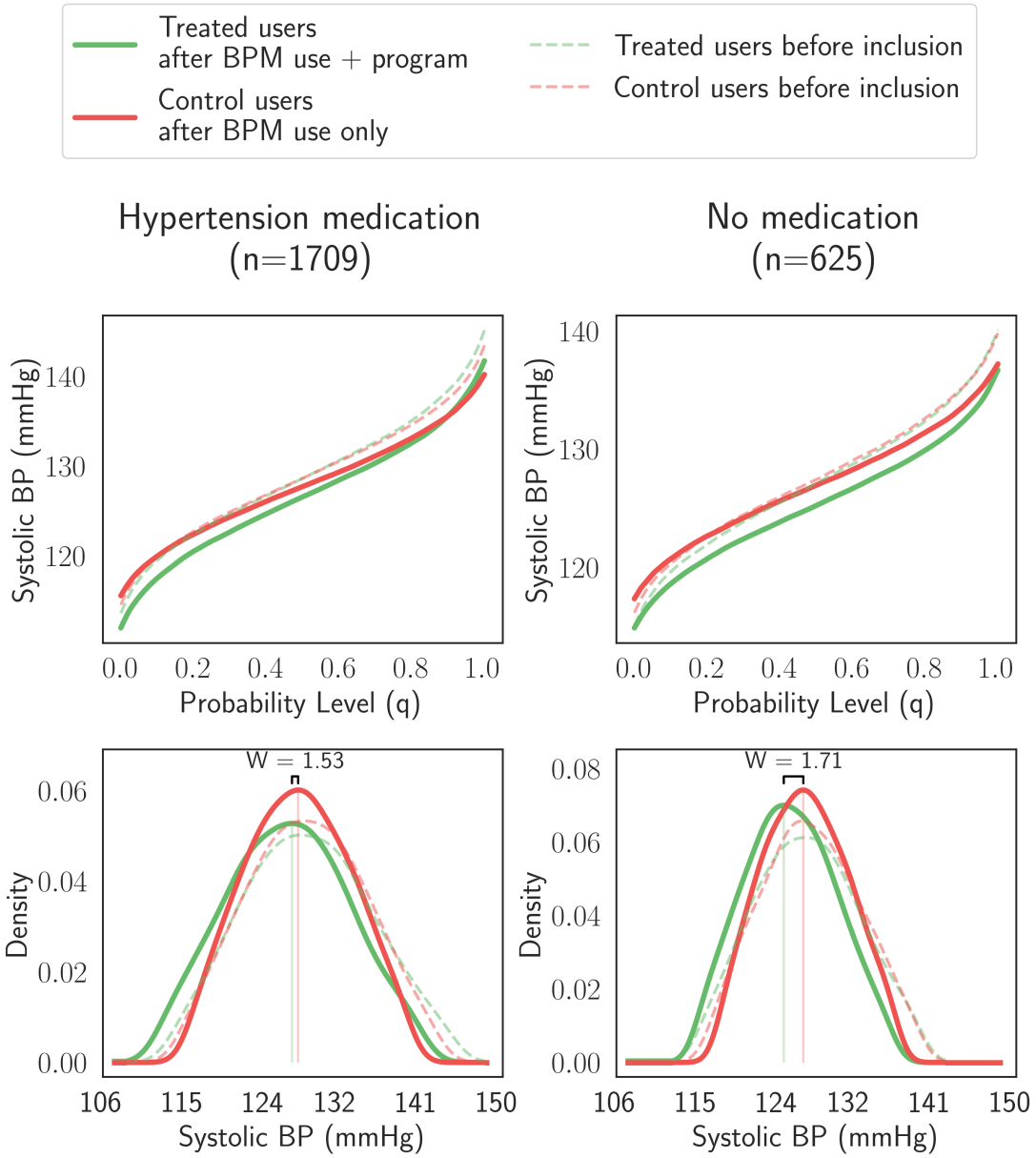


Figure 4: Fonction quantiles et densités de pressions artérielles des utilisateurs du programme (en vert) et des contrôles (en rouge) après matching, pour les sous-populations avec et sans prise de médicaments antihypertenseurs.

Génération hiérarchique de données de comptage avec PLN-Tree: augmentation de données en microbiologie

Alexandre Chaussard [1] & Anna Bonnet [1] & Sylvain Le Corff [1] & Harry Sokol [2]

[1] *Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, France, chaussard@lpsm.paris*

[2] *CHU Saint-Antoine Assistance publique Hôpitaux de Paris (AP-HP), Sorbonne Université, France*

Résumé. Le microbiote intestinal est un écosystème complexe décrit par des données de comptage obtenues par séquençage génétique. Constitué principalement de bactéries, l'équilibre de cet écosystème joue un rôle central dans la santé de l'hôte. Cependant, sa modélisation demeure un défi majeur en raison de ses dynamiques complexes et des propriétés spécifiques aux données de comptage compositionnelles. Récemment, [Chaussard et al., 2024] ont proposé une variante du modèle Poisson Log-Normal (PLN), standard pour l'analyse des données de comptage, qui intègre les relations taxonomiques entre les espèces. Ce modèle, nommé PLN-Tree, permet de générer des données artificielles qui reproduisent fidèlement des données de microbiote selon différentes métriques de diversité usuellement utilisées en écologie microbienne. Ces résultats suggèrent la pertinence du modèle PLN-tree pour réaliser une tâche d'augmentation de données. En effet, en apprenant le modèle sur des données microbiennes puis en générant de nouvelles données synthétiques, cette technique permet d'augmenter artificiellement la taille et la variété du jeu d'apprentissage, améliorant ainsi les performances et la robustesse des classifieurs. En particulier, nous nous focalisons sur des scénarios avec un faible nombre d'échantillons réels, en analysant des cohortes extraites de CuratedMetagenomics [Pasolli et al., 2017], et nous montrons que l'augmentation par PLN-Tree offre des gains compétitifs avec les techniques d'augmentation usuelles en microbiologie telles que Vanilla Mixup, Compositional Cutmix, ou encore PhyloMix [Gordon-Rodriguez et al., 2022, Jiang et al., 2025]. Par ailleurs, nous proposons une extension méthodologique de PLN-Tree en intégrant des covariables via des mécanismes d'attention. Nous comparons ainsi les approches conditionnelles et non-conditionnelles, soulignant l'impact de l'incorporation d'informations exogènes sur la qualité des générations et les performances de classification pour certaines maladies.

Keywords. Augmentation, classification, microbiote, données de comptage hiérarchiques

Abstract. The gut microbiome is a complex ecosystem described using count data obtained through genetic sequencing. While the balance of this ecosystem plays a crucial role in the host's health, its modeling remains a significant challenge due to its complexity and the unique properties of compositional count data. Recently, [Chaussard et al., 2024] introduced an extension of the standard Poisson Log-Normal (PLN) model—which is widely used for count data analysis—by incorporating taxonomic relationships among species. This

model, called PLN-Tree, has demonstrated improved data generation quality, thus motivating its application to data augmentation tasks. By generating synthetic microbiome count data, this approach can artificially expand both the size and diversity of training datasets, possibly yielding improved performances and robustness of the considered classifiers. In particular, we focus on the low sample size scenario, as mostly observed in real-world applications, leveraging cohorts from CuratedMetagenomics [Pasolli et al., 2017]. In particular, we show that augmentation using PLN-Tree is competitive with standard augmentation methods such as Vanilla Mixup, Compositional Cutmix, or PhyloMix [Gordon-Rodriguez et al., 2022, Jiang et al., 2025]. Additionally, we propose a methodological extension to PLN-Tree by incorporating covariates using attention mechanisms. This allows us to compare conditional and unconditional approaches, highlighting the impact of integrating exogenous information on data generation quality for various diseases.

Keywords. Data augmentation, classification, microbiome, hierarchical count data

1 Introduction

Les données de comptages apparaissent dans des domaines variés, allant de l'écologie aux sciences sociales en passant par l'actuariat. En particulier, on retrouve ce type de distribution dans l'étude du microbiote intestinal, qui est l'ensemble des micro-organismes qui peuplent notre tube digestif. L'équilibre du microbiote joue un rôle critique sur la santé de son hôte, et de nombreux liens entre certaines maladies et sa composition ont été établis [Kumar et al., 2024, Amos et al., 2021, Julien et al., 2022], suggérant son utilisation comme un biomarqueur potentiel. Cependant, la modélisation de la composition microbienne reste difficile, et peu de modèles statistiques semblent reproduire fidèlement ses caractéristiques. Récemment, [Chaussard et al., 2024] a introduit une variante du modèle Poisson Log-Normal (PLN) permettant d'incorporer les relations hiérarchiques dans la modélisation des écosystèmes microbiens. Cette approche générative nommée PLN-Tree permet notamment de capturer des propriétés d'α-diversité et de β-diversité qui caractérisent les écosystèmes et sont régulièrement utilisées pour décrire des états de santé [Zheng et al., 2024, Kers and Saccenti, 2022]. Cette observation clé assure la conservation des propriétés biologiques des données après augmentation par PLN-Tree, contrairement aux approches PLN, SPiEC-Easi qui ne peuvent donc être utilisées pour des tâches d'augmentation biologiquement interprétable, comme observé pour MB-GAN [Rong et al., 2021, Przymus et al., 2025].

Dans ces travaux, nous proposons donc d'appliquer le modèle PLN-Tree à des tâches d'augmentation de données pour la classification. En augmentant artificiellement la taille et la variété d'un jeu d'apprentissage, des données générées sous PLN-Tree peuvent permettre d'augmenter les performances et la robustesse des classifieurs, notamment les forêts aléatoires qui établissent le standard en microbiologie [Julien et al., 2022]. En raison des cohortes de petites tailles généralement observées, nous nous intéressons à des scénarios de faibles nombres d'échantillons réels de microbiotes ($n = 60$) issus de CuratedMetagenomics [Pasolli et al., 2017] pour apprendre PLN-Tree, illustrant l'intérêt pratique du modèle. L'augmentation est réalisée pour différentes maladies, illustrant des gains de performances comparables ou supérieur à ceux réalisés avec des méthodes d'augmentation clas-

siques en microbiologie telles que Vanilla Mixup, Compositional Cutmix, ou encore PhyloMix [Gordon-Rodriguez et al., 2022, Jiang et al., 2025].

Par ailleurs, nous étendons la contribution méthodologique de [Chaussard et al., 2024] en intégrant des informations exogènes (covariables) à l'aide de mécanismes d'attention. En particulier, on s'intéresse à l'intégration d'informations géographiques sur les patients, qui sont contenues dans les jeux de données et dont on sait qu'elles ont un effet important sur la composition microbienne [Gupta et al., 2017]. L'impact des covariables est illustrée par un benchmark des qualités génératives du modèle face à PLN-Tree sans covariables.

2 Le modèle PLN-Tree

Modèles PLN. Pour des données tabulaires de comptages $\mathbf{X} \in \mathbb{R}^{n \times d}$, le modèle Poisson Log-Normal (PLN), introduit par [Aitchison and Ho, 1989], est décrit pour une variable latente associée $\mathbf{Z} \in \mathbb{R}^{n \times d}$ par

$$\begin{aligned} \mathbf{Z} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ , \\ \mathbf{X} &\sim \mathcal{P}(\mathrm{e}^{\mathbf{Z}}) \ , \end{aligned}$$

où $\boldsymbol{\mu} \in \mathbb{R}^d$ et $\boldsymbol{\Sigma} \in \mathcal{S}_+^d$, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ représente les paramètres du modèle. La matrice $\boldsymbol{\Sigma}$ caractérise les interactions entre les entités du système et $\boldsymbol{\mu}$ représente des effets fixes liés au milieu. L'inférence du modèle, développée dans [Chiquet et al., 2019], peut être réalisée par une approche variationnelle en maximisant la quantité sous-optimale de l'Evidence Lower BOund (ELBO) [Kingma et al., 2019] :

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \mathbb{E}_{q_{\boldsymbol{\varphi}}}[\log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})] - \mathrm{KL}[q_{\boldsymbol{\varphi}}(\cdot|\mathbf{X}), p_{\mathbf{Z},\boldsymbol{\theta}}(\cdot)] \ . \tag{1}$$

L'approximation variationnelle proposée par [Chiquet et al., 2021] repose sur l'approximation champ-moyen Gaussienne traditionnelle :

$$q_{\boldsymbol{\varphi}}(\mathbf{Z}_i|\mathbf{X}_i) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{Z}_i; \mathbf{m}_i, \mathbf{S}_i) \ ,$$

où $\mathcal{N}(\cdot; \mathbf{m}_i, \mathbf{S}_i)$ est la densité de la loi Gaussienne de moyenne $\mathbf{m}_i \in \mathbb{R}^d$ et de variance $\mathbf{S}_i = \mathrm{diag}(\mathbf{s}_i)$ avec $\mathbf{s}_i \in \mathbb{R}^d$ et où $\boldsymbol{\varphi}$ représente tous les paramètres variationnels. L'apprentissage des paramètres repose sur une optimisation alternée détaillée dans [Chiquet et al., 2021].

Le modèle PLN propose donc un cadre adapté aux données multivariées de comptage, satisfaisant des propriétés élémentaires de surdispersion des comptes et d'identifiabilité des paramètres [Chaussard et al., 2024]. Ses avantages ont favorisé le développement d'un large spectre d'outils d'analyse des comptes tels que l'ACP [Chiquet et al., 2018], l'analyse discriminante [Chiquet et al., 2021] et l'inférence de réseau [Chiquet et al., 2019].

Intégration de hiérarchies avec PLN-Tree. Parmi les extensions du modèle PLN, le modèle PLN-Tree introduit par [Chaussard et al., 2024] permet l'intégration d'information

hiérarchique liant les entités d'un système. La structure attendue des comptes n'est plus tabulaire, mais repose sur un arbre couvrant comme illustré en Figure 1.

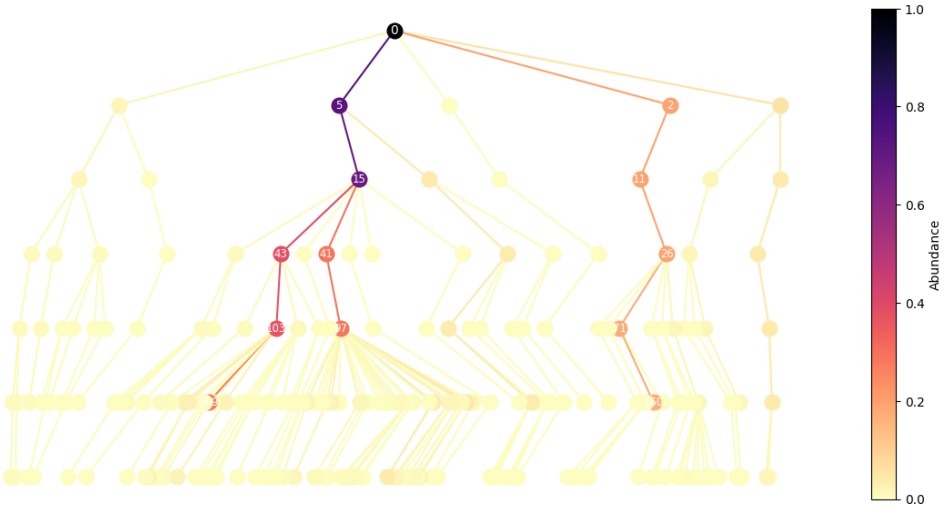


Figure 1: Exemple d'échantillon d'abondance de taxons (coloré selon l'abondance relative)

En particulier, la parenté des noeuds implique une contrainte compositionnelle entre parent et enfants, de sorte que si ℓ indique un niveau de l'arbre, k un noeud associé à ce niveau, $\mathcal{C}_k^\ell$ les indices des noeuds enfants associés au parent $k \leqslant K_\ell$ à la couche $\ell < L$, alors la compositionnalité hiérarchique des comptes s'écrit

$$\mathrm{X}_k^\ell = \sum_{j \in \mathcal{C}_k^\ell} \mathrm{X}_j^{\ell+1} \,. \tag{2}$$

Le modèle PLN-Tree assure l'intégrité des données hiérarchiques en intégrant la contrainte (2) dans une dynamique Markovienne Poisson log-normal dans l'espace des comptes, tandis que la dynamique latente Markovienne Gaussienne se propage sans contrainte. Ainsi, le modèle PLN-Tree permet d'inclure l'information hiérarchique sans restreindre les interactions à la structure d'arbre, assurant donc une flexibilité vis-à-vis de la hiérarchie. Si l'on note $\check{\mathbf{X}}_k^\ell$ les enfants associés au noeud $k \leqslant K_\ell$ à la couche $\ell < L$, la dynamique de PLN-Tree s'écrit :

$$\begin{aligned}
&\mathbf{Z}^1 \sim \mathcal{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right) \,, \quad \mathbf{Z}^{\ell+1} \mid \mathbf{Z}^\ell \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}_{\ell+1}}(\mathbf{Z}^\ell), \boldsymbol{\Sigma}_{\ell+1}(\mathbf{Z}^\ell)\right) \,, \\
&\mathbf{X}^1 \mid \mathbf{Z}^1 \sim \mathcal{P}(\mathrm{e}^{\mathbf{Z}^1}) \,, \quad \check{\mathbf{X}}_k^\ell \mid \mathrm{X}_k^\ell, \check{\mathbf{Z}}_k^\ell \sim \mathcal{M}\left(\mathrm{X}_k^\ell, \sigma(\check{\mathbf{Z}}_k^\ell)\right) \,,
\end{aligned}$$

où $\boldsymbol{\mu}_1 \in \mathbb{R}^{K_1}, \boldsymbol{\Sigma}_1 \in \mathcal{S}_+^{K_1}$ et pour tout $\ell < L, \boldsymbol{\mu}_{\boldsymbol{\theta}_{\ell+1}}$ et $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\ell+1}}$ sont des fonctions paramétrées par $\boldsymbol{\theta}_{\ell+1}$ à valeurs respectivement dans $\mathbb{R}^{K_{\ell+1}}$ et $\mathcal{S}_+^{K_{\ell+1}}$. L'apprentissage des paramètres du modèle $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_L)$ est réalisé par une approche variationnelle s'appuyant sur des travaux théoriques récents en inférence amortie backward [Chagneux et al., 2022, Campbell et al., 2021] détaillée dans [Chaussard et al., 2024].

Cependant, la dépendance hiérarchique introduit des difficultés d'interprétabilité des paramètres en toute généralité. Pour répondre à ce problème, [Chaussard et al., 2024] présente

une classe d'identifiabilité du modèle assurant l'interprétabilité des réseaux appris et présentant des conséquences pratiques pour la classification. Par ailleurs, les expériences menées dans ces travaux mettent en évidence l'intérêt spécifique de la taxonomie pour la génération de données microbiennes, motivant l'application du modèle PLN-Tree à des problématiques d'augmentation de données.

3 Augmentation de données microbiennes avec PLN-Tree

Suite aux progrès réalisés en génération de données microbiennes dans [Chaussard et al., 2024], mesurés sous des critères d'α-diversité et de β-diversité, on s'intéresse à la génération de microbiotes artificiels pour augmenter la taille du jeu d'apprentissage pour différents problèmes de classification de maladies. En enrichissant le jeu d'apprentissage, cette technique vise à améliorer les performances et la robustesse des classifieurs, notamment lorsque le nombre d'échantillons est limité (ici $n = 60$).

Méthodes d'augmentation. Pour les données de microbiote, [Gordon-Rodriguez et al., 2022] proposent des méthodes exploitant la compositionalité des données, notamment Aitchison Mixup et Compositional Cutmix. La méthode historique Vanilla Mixup [Zhang et al., 2017] est également valide dans ce cadre comme rappelée par les auteurs. En utilisant l'information phylogénétique, PhyloMix [Jiang et al., 2025] propose également une variante du Compositional Cutmix avec des performances supérieures. Suivant le protocole de [Jiang et al., 2025] et les résultats illustrés, nous nous comparerons donc aux approches Vanilla Mixup, Compositional Cutmix et PhyloMix qui semblent compétitives dans notre cadre.

Augmentation sans covariables. Dans un premier temps, nous entraînons un modèle génératif PLN-Tree sur des données de microbiotes issues de diverses études portant sur différentes maladies, regroupées la base de données CuratedMetagenomics [Pasolli et al., 2017]. Nous nous focalisons sur des cohortes de taille réduite (60 échantillons équilibrés), où les performances des forêts aléatoires sont sous-critiques dans la distinction recherchée. Le modèle PLN-Tree est d'abord appris sur ces petites cohortes, puis utilisé pour augmenter le nombre d'échantillons de la cohorte. Cette procédure est réalisée sur divers scénarios de classification binaire équilibrés avec 60 échantillons : adénome / control, diabète de type 2 / control, mélanome / control, colitis / control, cancer colorectal /control, colitis / crohn, adénome / cancer colorectal. Pour chacun des scénarios, l'augmentation de données est réalisée en augmentant la taille de l'échantillon initial de 200%, comme effectué dans le protocole de PhyloMix [Jiang et al., 2025]. L'augmentation utilisant PLN-Tree permet d'améliorer les performances des forêts aléatoires dans la majorité des cas, indiquant des gains comparables ou supérieurs à ceux réalisés avec des approches non-paramétriques. Les résultats sont illustrés en Figure 2.

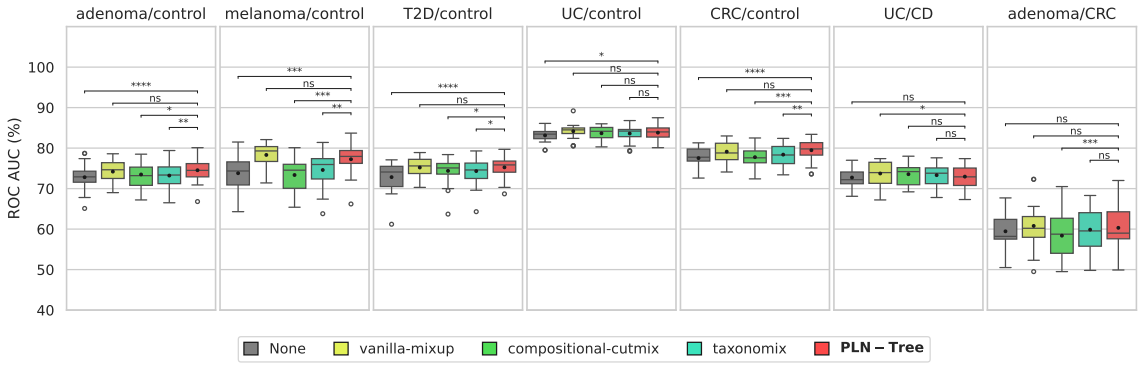


Figure 2: Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) pour différents scénarios de classification de microbiotes utilisant les forêts aléatoires par défaut de scikit-learn. La classification est répétée sur 30 folds. La significativité des résultats est évaluée entre chaque méthode par un t-test (ns: non significatif, *: $p \leqslant 0.5$, **: $p \leqslant 0.01$, ***: $p \leqslant 0.001$, ****: $p \leqslant 0.0001$). "Taxonomix" correspond à la stratégie PhyloMix utilisant la taxonomie.

Augmentation conditionnelle. Dans [Chaussard et al., 2024], l'introduction de covariables (information exogène) dans le modèle n'est pas abordée. Nous proposons donc une architecture basée sur des mécanismes d'attention pour incorporer des covariables dans le modèle. Nous étudions notamment l'impact de la localisation des patients sur la qualité des générations, compte tenu des effets critiques identifiés dans [Gupta et al., 2017].

References

[Aitchison and Ho, 1989] Aitchison, J. and Ho, C. (1989). The multivariate poisson-log normal distribution. *Biometrika*, 76(4):643–653.

[Amos et al., 2021] Amos, G. C., Sergaki, C., Logan, A., Iriarte, R., Bannaga, A., Chandrapalan, S., Wellington, E. M., Rijpkema, S., and Arasaradnam, R. P. (2021). Exploring how microbiome signatures change across inflammatory bowel disease conditions and disease locations. *Scientific Reports*, 11(1):18699.

[Campbell et al., 2021] Campbell, A., Shi, Y., Rainforth, T., and Doucet, A. (2021). Online variational filtering and parameter learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 18633–18645.

[Chagneux et al., 2022] Chagneux, M., Gassiat, É., Gloaguen, P., and Corff, S. L. (2022). Amortized backward variational inference in nonlinear state-space models. *arXiv preprint arXiv:2206.00319*.

[Chaussard et al., 2024] Chaussard, A., Bonnet, A., Gassiat, E., and Corff, S. L. (2024). Tree-based variational inference for poisson log-normal models. *arXiv preprint arXiv:2406.17361*.

[Chiquet et al., 2018] Chiquet, J., Mariadassou, M., and Robin, S. (2018). Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, 12(4):2674 – 2698.

[Chiquet et al., 2021] Chiquet, J., Mariadassou, M., and Robin, S. (2021). The poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, 9:588292.

[Chiquet et al., 2019] Chiquet, J., Robin, S., and Mariadassou, M. (2019). Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning*, pages 1162–1171. PMLR.

[Gordon-Rodriguez et al., 2022] Gordon-Rodriguez, E., Quinn, T., and Cunningham, J. P. (2022). Data augmentation for compositional data: Advancing predictive models of the microbiome. *Advances in Neural Information Processing Systems*, 35:20551–20565.

[Gupta et al., 2017] Gupta, V. K., Paul, S., and Dutta, C. (2017). Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Frontiers in microbiology*, 8:1162.

[Jiang et al., 2025] Jiang, Y., Liao, D., Zhu, Q., and Lu, Y. Y. (2025). Phylomix: Enhancing microbiome-trait association prediction through phylogeny-mixing augmentation. *Bioinformatics*, page btaf014.

[Julien et al., 2022] Julien, C., Anakok, E., Treton, X., Nachury, M., Nancey, S., Buisson, A., Fumery, M., Filippi, J., Maggiori, L., Panis, Y., et al. (2022). Impact of the ileal microbiota on surgical site infections in crohn's disease: a nationwide prospective cohort. *Journal of Crohn's and Colitis*, 16(8):1211–1221.

[Kers and Saccenti, 2022] Kers, J. G. and Saccenti, E. (2022). The power of microbiome studies: some considerations on which alpha and beta metrics to use and how to report results. *Frontiers in microbiology*, 12:796025.

[Kingma et al., 2019] Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

[Kumar et al., 2024] Kumar, B., Lorusso, E., Fosso, B., and Pesole, G. (2024). A comprehensive overview of microbiome data in the light of machine learning applications: categorization, accessibility, and future directions. *Frontiers in Microbiology*, 15:1343572.

[Pasolli et al., 2017] Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., et al. (2017). Accessible, curated metagenomic data through experimenthub. *Nature methods*, 14(11):1023–1024.

[Przymus et al., 2025] Przymus, P., Rykaczewski, K., Martín-Segura, A., Truu, J., Carrillo De Santa Pau, E., Kolev, M., Naskinova, I., Gruca, A., Sampri, A., Frohme, M., et al. (2025). Deep learning in microbiome analysis: a comprehensive review of neural network models. *Frontiers in Microbiology*, 15:1516667.

[Rong et al., 2021] Rong, R., Jiang, S., Xu, L., Xiao, G., Xie, Y., Liu, D. J., Li, Q., and Zhan, X. (2021). Mb-gan: microbiome simulation via generative adversarial network. *GigaScience*, 10(2):giab005.

[Zhang et al., 2017] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *ICLR 2018*.

[Zheng et al., 2024] Zheng, J., Sun, Q., Zhang, M., Liu, C., Su, Q., Zhang, L., Xu, Z., Lu, W., Ching, J., Tang, W., et al. (2024). Noninvasive, microbiome-based diagnosis of inflammatory bowel disease. *Nature Medicine*, 30(12):3555–3567.

A Model-Based Clustering approach for toxicity assessment using Cell Painting data

Mariam Grigoryan [1] & Vincent Vandewalle [2] & David Rouquié [3] & Yoo Jin Lee [4]

[1] *University Côte d'Azur, Laboratoire J.A. Dieudonné, Inria (équipe MAASAI), CNRS, France, mariam.grigoryan@univ-cotedazur.fr*
[2] *University Côte d'Azur, Laboratoire J.A. Dieudonné, Inria (équipe MAASAI), CNRS, France, vincent.vandewalle@univ-cotedazur.fr*
[3] *Bayer Crop Science, Toxicology Data Science Department, France, david.rouquie@bayer.com*
[4] *Bayer Crop Science, Toxicology Data Science Department, France, yoojin.lee@bayer.com*

Résumé. Le *Cell Painting* est une méthode d'imagerie à haut contenu qui permet d'analyser plusieurs compartiments cellulaires simultanément, offrant ainsi une vue d'ensemble de l'impact des produits chimiques sur la morphologie des cellules et la structure des organites. Dans ce travail nous étudions ces changements morphologiques dans un cadre dose-réponse où l'objectif est d'identifier la première concentration à laquelle des changements morphologiques significatifs sont observés. Nous proposons de résoudre le problème dose-réponse à haute dimension par une approche en deux étapes. Tout d'abord, un *clustering* des cellules dans chaque puits des plaques à essai est effectué, ce qui permet de résumer chaque puits par un ensemble de paramètres de faible dimension. Ensuite, une approche de tests séquentiels est réalisée pour trouver la concentration minimale à laquelle un composé chimique induit une différence significative sur les caractéristiques des cellules par rapport aux puits de contrôle négatifs. Le *clustering* est effectué en ajustant un modèle de mélange gaussien dans lequel les paramètres spécifiques aux classes sont supposés être les mêmes dans tous les puits, mais les proportions des classes sont supposées être libres dans chaque puits. Il est ainsi possible de résumer l'information de chaque puits par les proportions des différentes classes dans ce puit. La deuxième étape consiste à tester si le vecteur des proportions des classes pour un composé chimique considéré, à une concentration donnée, a une distribution significativement différente du vecteur des proportions pour les contrôles négatifs. À cette fin, nous utilisons un test de permutation non paramétrique et une approche de tests séquentiels (tests pour des valeurs croissantes de la concentration du composé). L'approche est appliquée à des données réelles de Cell Painting montrant sa capacité à détecter des changements de distribution pour des concentrations variables des composés.

Mots-clés. Clustering à base de modèles, tests séquentiels, évaluation de toxicité, modèles dose-réponse

Abstract. Cell Painting is a high-content imaging approach that allows for the analysis of several cellular compartments, offering a comprehensive view of how chemicals impact cell morphology and organelle structure. In this work, these morphological changes are studied in a dose-response setting with the goal of identifying the Point of Departure (POD),

the first concentration at which significant morphological changes are observed. The dose-response problem is solved by a two step approach. First, a clustering of the cells within each well of the assay plates is performed thus allowing to summarize each well by a low dimensional set of parameters. Second, a sequential testing approach is performed for finding the minimal concentration at which a chemical compound induces a significant difference on cells characteristics compared with the negative control wells. The clustering is performed by fitting a Gaussian mixture model where class-specific parameters are assumed to be the same across wells but class proportions are assumed to be free in each well. Thus, it makes it possible to summarize the information of each well by its specific class proportions. The second step consists of testing if the vector of class proportions for a considered chemical compound at a given concentration has a distribution significantly different from the vector of proportions for the negative controls. To this aim we use a non-parametric permutation test, and a sequential testing approach (tests for increasing values of the compound concentration). The approach is applied to Cell Painting data showing its ability to detect distribution shifts for varying compound concentrations.

Keywords. Model-based clustering, sequential testing, toxicity assessment, dose-response models

1 Introduction

Cell-based assays like Cell Painting allows for chemical toxicity screening, offering a potential alternative to animal testing. This high-content imaging technique captures cellular morphology that can be used to investigate dose-dependent responses to chemicals (Bray et al. 2016). A key goal in toxicology is determining the Point of Departure (POD), the lowest concentration inducing measurable effects. Traditional methods, such as concentration-response modeling consider well-level aggregated features (Phillips et al. 2019), which may not fully capture the complexity of cellular responses. To overcome these limitations, we propose a single-cell modeling approach using Gaussian Mixture Models (GMMs) (McLachlan and Peel 2000) to analyze high-dimensional Cell Painting data. Our method identifies homogeneous subpopulations of morphological profiles and tracks their dose-dependent shifts. This approach offers biologically meaningful POD estimates by directly modeling single-cell heterogeneity, improving toxicity assessment in high-dimensional datasets.

2 Model-based clustering of cells with well-free class proportions

Model presentation

Let $\mathbf{x}_{ij}$ ($\mathbf{x}_{ij}$ in $\mathbb{R}^d$) be the feature vector of cell j ($j \in \{1, \ldots, m_i\}$) in well i ($i \in \{1, \ldots, n\}$) exposed on substance s_i at concentration c_i. We assume that $\mathbf{x}_{ij}$ follows a GMM where the mixing proportions $\boldsymbol{\pi}_i = (\pi_{i1}, ..., \pi_{ik})$ depend on the well. Thus the probability distribution (pdf) of $\mathbf{x}_{ij}$ is $p(\mathbf{x}_{ij} \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_{ik} \mathcal{N}(\mathbf{x}_{ij} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ where K is the number of Gaussian components (clusters) and $\mathcal{N}(\mathbf{x}_{ij} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the pdf of the multivariate Gaussian distribution

with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ which is assumed to be diagonal, $\boldsymbol{\theta}$ groups all the model parameters.

Parameters estimation and model selection

The model parameters are estimated by maximum likelihood using the EM algorithm, only needing to adapt the standard EM algorithm to varying class proportions. The optimal number of clusters K, is determined using the Bayesian Information Criterion (BIC).

3 A statistical test for dose-response assessment

For a given compound the goal is to find the lowest concentration at which the distribution of morphological features across cells is different from those of negative controls. We propose to build a non-parametric test based on the distribution of $\hat{\boldsymbol{\pi}}_i$ for each compound and each concentration, then use a sequential approach for multiple test correction in order to determine the lowest concentration at which the effect is observed.

Non-parametric permutation test on class proportions

The experimental design considers several replicates plates containing the different substances at different concentrations and also the negative controls. We consider for each replicate plate r the difference between the estimated average negative control proportions (denoted by $\hat{\boldsymbol{\pi}}^{0r}$) and the average of the considered compound s at a given concentration c class proportions (denoted by $\hat{\boldsymbol{\pi}}^{scr}$) as test statistic which is denoted by $T_{scr,obs} = \|\hat{\boldsymbol{\pi}}^{0r} - \hat{\boldsymbol{\pi}}^{scr}\|_1$. Assuming the same distribution between the compound s at concentration c and the negative control for replicate r we can define a random permutation p of the vector composed with $\hat{\boldsymbol{\pi}}_i$ related to negative controls and compound s at concentration c in plate r and obtain the permutation statistic $T^p_{scr,perm}$ ($p \in \{1, \ldots, P\}$) to deduce the related p-value as $p\text{-value}_{scr} = \frac{1}{P}\sum_{p=1}^{P} \mathbb{1}(T^p_{\text{perm},scr} > T_{\text{obs},scr})$ where P is the number of permutations considered.

In fact it is more relevant to consider a global test on all replicates rather than a test for each replicate thus we consider the following test statistic for compound s at concentration c, $T_{sc,\text{obs}} = \sum_{r=1}^{R} T_{\text{obs},scr}$ then still determine the distribution of this test statistic under H_0 by random permutation of the corresponding $\hat{\boldsymbol{\pi}}$ related to each replicate.

Sequential test for increasing compound concentrations

In multiple hypothesis testing, particularly when hypotheses follow a known order, it is crucial to control the False Discovery Rate (FDR) while maintaining statistical power. Traditional methods, such as the Bonferroni correction tend to be overly conservative, reducing the ability to detect true effects. The Adaptive SeqStep (AS) procedure addresses this by sequentially testing hypotheses in a predefined order while dynamically adjusting the rejection threshold (Lei and Fithian 2016). This approach optimizes the trade-off between detecting true signals and controlling false discoveries. Unlike fixed-threshold methods, AS estimates the proportion of null hypotheses adaptively, ensuring that more hypotheses can be rejected while keeping the FDR under control. In our case, where higher concentrations are more likely to induce

significant effects, AS prioritizes these hypotheses. The procedure stops once the estimated false discovery proportion exceeds a predefined threshold, preventing excessive false positives while maximizing meaningful discoveries. This makes AS particularly well-suited for structured experimental designs, such as dose-response studies.

4 Experiments on Cell Painting data

Presentations of the data

The data used in this study were obtained from Bayer Crop Science using the Cell Painting technique. Approximately 2,000 U2OS cells were seeded into each well of a 384-well plate. Each well contains cells treated with different chemical compounds.

For this study, we focus on one compound, resulting in data from eight wells with different concentrations. In addition to these, we include data from six wells, which are negative controls treated with Dimethyl Sulfoxide (DMSO).

The experiment was performed with four replicates. After the experiment, each well of every plate was photographed, capturing the cells using a microscope. These images were then segmented using Cell Profiler (Institute 2021), an open-source image analysis software. This tool enabled the extraction of quantitative data on cells, such as their shape, size, density, and distribution, which we refer to as morphological profiles.

Results

We start by reducing the data by selecting only the jump features, resulting in a total of 560 features.Those features refer to a curated subset of features identified by the Joint Undertaking for Morphological Profiling (JUMP) Cell Painting consortium (MIT and Harvard 2024). To determine the optimal number of clusters, we use the BIC criterion as outlined earlier. Below, we present the results obtained for cluster distribution ($K = 2$):

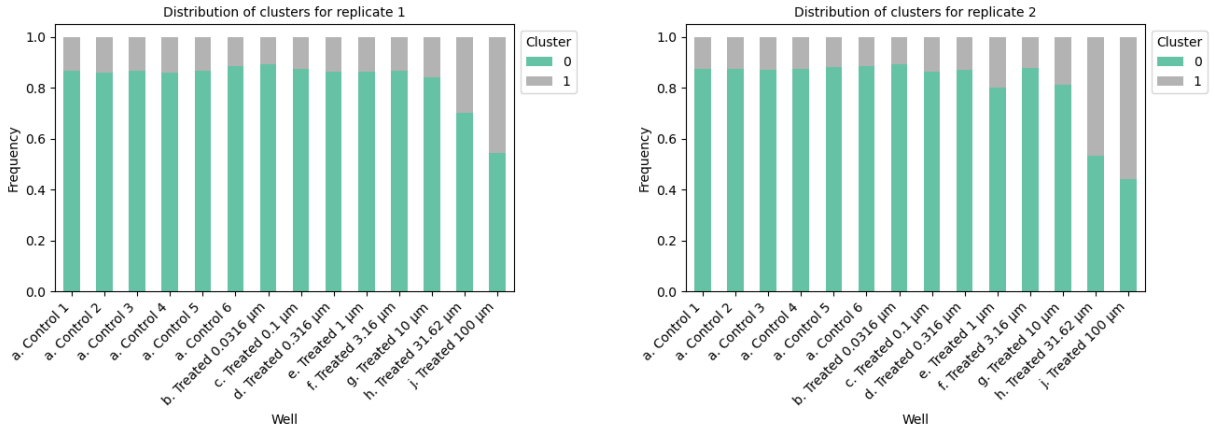


Figure 1: Distribution of clusters for replicate 1 and 2

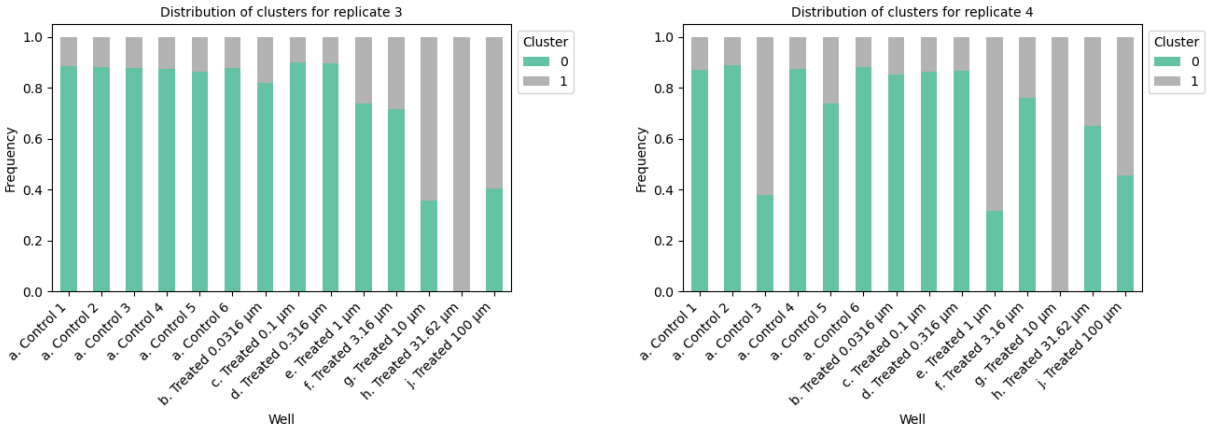


Figure 2: Distribution of clusters for replicate 3 et 4

Figures 1 and 2 present the distribution of the clusters across the four replicates. First on the left, we observe the negative controls. Next, we move on to the treated wells presented in increasing order of concentration, from the lowest to the highest.

It is noteworthy that replicates 3 and 4 show significant variability, especially replicate 4. However, there is a clear similarity between the cluster distribution of the negative controls and that of our treated compound at low concentration. In contrast, at high concentration, we observe an increase in variability in the cluster distribution.

To confirm these initial observations, we perform permutation tests to validate our hypotheses and investigate the impact of concentration on cluster distributions. Following these tests, we apply the AS procedure to control the FDR while accounting for multiple testing.

Replicate	**Concentration** (μM)							
	0.0316	**0.1**	**0.316**	**1**	**3.16**	**10**	**31.6**	**100**
1	✓	✓	✓	✓	×	×	×	×
2	✓	✓	×	×	×	×	×	×
3	✓	×	×	×	×	×	×	×
4	✓	✓	✓	✓	✓	✓	×	×
Global Test	✓	✓	✓	✓	×	×	×	×

Table 1: Results of permutation tests for different replicates and concentrations.

Table 1 presents the obtained results when performing the test on each replicate separately or when performing the global test based on the aggregated statistic. A rejected null hypothesis (×) indicates that the treated wells show significant differences in cluster proportions compared to those in negative controls, suggesting a possible effect of the treatment at that concentration. On the other hand, a retained null hypothesis (✓) means no significant difference are detected at that concentration.

The Adaptive SeqStep method controls the FDP by ensuring that the proportion of false discoveries among the rejected hypotheses remains below a predefined level q. This threshold

q represents the maximum tolerated proportion of false discoveries in the selected set. In our case we set $q = 0.1$.

To conclude, the concentrations of 31.6 and 100 μM unanimously showed significant differences across all replicates. However, due to the evident technical variability in the dataset caution is required when interpreting these results, particularly from 3.16 μM for this specific compound. To ensure the robustness of our approach further analyses should be conducted on a broader range of compounds.

5 Conclusion and perspectives

In this study, we propose a model-based clustering approach for dose-response assessment. This method provides biologically meaningful estimates of the POD, capturing dose-dependent changes in cellular morphology and structure. Our experiments on Cell Painting data demonstrate the effectiveness of this approach in detecting concentration-dependent distribution shifts, offering an efficient tool for chemical toxicity assessment.

While this approach has shown promising results it has certain limitations. First and foremost, inter-replicate variability must be carefully considered to ensure the robustness of the results. Then, the findings are based on a single cell line which may not fully capture the complexity of an entire organism. Moreover, converting in vitro concentrations to real life exposure levels requires inverse dosimetry which complicates extrapolation. Finally, morphological changes do not always correlate with toxicity, necessitating complementary analyses to confirm adverse effects.

Future work will address these limitations by extending the analysis to multiple cell types and integrating complementary data sources, such as gene expression profiles, to provide a more comprehensive view of chemical-induced effects.

References

Bray, Mark-Anthony et al. (2016). "Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes". In: *Nature Protocols* 11.9, pp. 1757–1774. ISSN: 1750-2799.

Institute, Broad (2021). "Cell Profiler". In: URL: https://cellprofiler.org.

Lei, Lihua and William Fithian (2016). "Power of Ordered Hypothesis Testing". In: Proceedings of Machine Learning Research 48. Ed. by Maria Florina Balcan and Kilian Q. Weinberger, pp. 2924–2932.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. New York: Wiley-Interscience.

MIT, Broad Institute of and Harvard (2024). *JUMP-Cell Painting*. https://jump-cellpainting.broadinstitute.org. Accessed: 2024-12-11.

Phillips, Jason R et al. (2019). "BMDExpress 2: enhanced transcriptomic dose-response analysis workflow". en. In: *Bioinformatics* 35.10. Ed. by Alfonso Valencia, pp. 1780–1782. ISSN: 1367-4803, 1367-4811. (Visited on 02/02/2024).

ToMATo clustering algorithm for neuronal spike sorting

Ségolen Geffray [1]& Louise Martineau [1] & Christophe Pouzat [1]

[1] *Institut de Recherche Mathématique Avancée Strasbourg, UMR 7501 Université de Strasbourg et CNRS, 7 rue René-Descartes, 67000 Strasbourg, France,*
{*segolen.geffray, louise.martineau, christophe.pouzat*}*@math.unistra.fr*

Résumé. Les enregistrements extracellulaires avec matrices multi-éléctrode sont un outil basique utilisé en neuroscience expérimentale. Les données brutes collectées sont des séquences d'impulsion ou potentiels d'action qui ont été générés par plusieurs neurones. Pour trouver le nombre de neurones et attribuer chaque potentiel d'action à un neurone, une étape de prétraitement appelée tri des potentiels d'action est nécessaire. C'est un procédé semi-automatique qui comprend beaucoup d'étapes, la dernière étant un problème de partitionnement en grande dimension. La procédure nécessite donc une réduction de la dimension, qui est sensible à la présence de superpositions d'événements (comparable à des valeurs aberrantes), menant à de mauvais résultats de partitionnement. Les neuroscientifiques sont alors généralement amenés à enlever ces superpositions, une procédure délicate. L'utilisation de l'algorithme de partitionnement ToMATo aide à simplifier et alléger tout ce procédé de tri des potentiels d'action. ToMATo (Topological Mode Analysis Tool) est une méthode de partitionnement provenant du domaine de l'analyse topologique des données. C'est un algorithme de *mode-seeking*, qui cherche des pics dans la densité f des observations, et attribue les observations tombant sous le pic au même cluster. Cet algorithme utilise les estimations de la densité seulement aux points des données et effectue un *mode-seeking* basé sur les graphes plutôt que de rechercher directement les pics de $\hat{f}$, ce qui rend la méthode utilisable même en grande dimension. En outre, avec la théorie de la persistance une notion de proéminence de pics est définie et les clusters déterminés par le *mode-seeking* sont fusionnés de sorte que les clusters finaux correspondent seulement aux pics proéminents de la vraie densité f, retrouvant ainsi le bon nombre de clusters. On montre que ToMATo peut être appliqué facilement pour le tri des potentiels d'action, permettant de réduire le nombre d'étapes généralement nécessaires dans cette procédure. En effet, puisque cet algorithme peut être appliqué directement sur des données de grandes dimension (dans notre application les données sont de dimension 180), la réduction de dimension qui précède le partionnement n'est plus nécessaire, ce qui fait également disparaitre la sensibilité aux superpositions. Notre cas montre les performances spectaculaires de cette approche.

Mots-clés. Tri des potentiels d'action, partitionnement, homologie persistante

Abstract. Extracellular recordings with multi-electrode arrays is one of the basic tools used in experimental neuroscience. The collected raw data are sequences of impulses or spikes that are generated by several neurons. Finding out the number of neurons as well as assigning each spike to a given neuron is a required preprocessing step called spike sorting. It is a semi-automatic process that involves many steps, the last one being a clustering problem in high dimension. It is therefore accompanied by a dimension reduction, which is sensitive to the presence of event superposition (akin to outliers) leading to poor clustering results. Neuroscientists are then usually led to remove these superpositions, a delicate procedure. The use of the ToMATo clustering algorithm helps to simplify and streamline this whole spike sorting procedure. ToMATo (Topological Mode Analysis Tool) is a clustering method coming from the field of topological data analysis. It is a mode-seeking algorithm, that looks for peaks in the observations density f and assigns observations falling under the same peak to the same cluster. This algorithm uses density estimate only at the data points and performs

a graph-based mode-seeking instead of peak localisation on $\hat{f}$, making it usable even in high dimension. Moreover with persistence theory a notion of peak prominence is defined and clusters found by mode-seeking are merged together so that the final clusters correspond only to prominent peaks of the true density f, recovering the correct number of clusters. We show that ToMATo can be easily applied for spike sorting, enabling to reduce the numbers of steps typically involved in this procedure. Indeed, since it can be applied directly to high dimensional data (in our application the data is 180-dimensional), there is no need for dimension reduction before clustering, making the sensitivity to superpositions disappear. Our use case demonstrates the spectacular performances of this approach.

Keywords. Neuronal spike sorting, clustering, persistent homology

1 L'algorithme de partitionnement ToMATo

L'algorithme ToMATo (Topological Mode Analysis Tool) est une méthode de partitionnement qui a été introduite en 2013 dans l'article [Chazal et al., 2013]. C'est un algorithme de *mode-seeking*, c'est à dire qu'il consiste à rechercher des pics dans la densité des observations, et à définir un cluster comme l'ensemble des points tombant sous un pic. La vraie densité f n'étant jamais connue, les algorithmes de *mode-seeking* cherchent alors des pics d'un estimateur $\hat{f}$. Or, le bruit peut induire un grand nombre de pics peu proéminents de $\hat{f}$ et ainsi conduire à un trop grand nombre de clusters. La théorie de la persistance permet de définir la proéminence des pics, et guarantit que les pics proéminents d'un estimateur $\hat{f}$ vont correspondre aux pics proéminents de la densité f. L'originalité de ToMATo est alors de rajouter une étape de fusion des clusters en fonction de la proéminence du pic associé, de sorte que les clusters finaux ne correspondent qu'à des pics importants de la densité et pas à des pics de $\hat{f}$ induits par du bruit. Expliquons ce principe sur un exemple : imaginons que la fonction représentée dans la Figure 1 corresponde à un estimateur d'une certaine densité de $\mathbb{R}$ dans $\mathbb{R}^+$. On a envie de penser qu'il n'y a que 2 pics dans la vraie densité f, et que les petits pics que l'on observe sont dus au bruit.

On applique l'homologie persistante sur $\hat{f}$ pour définir la proéminence de ses pics. Rappelons-en le principe : ici on va s'intéresser aux composantes connexes des sur-niveaux de $\hat{f}$, en faisant baisser le niveau de $+\infty$ à 0. On peut imaginer qu'on fait baisser le niveau de la mer et que l'on voit des îles apparaître. Dans la Figure 1 A, on représente les composantes connexes de $\{\hat{f} \geq \alpha\}$ pour une certaine valeur de α. Maintenant la valeur de α varie en décroissant : les sur-niveaux de $\hat{f}$ sont vides pour α trop grand, puis lorsque $\alpha = \hat{f}(p_1)$, on voit apparaître une composante connexe dans $\{\hat{f} \geq \alpha\}$, qui correspond au plus grand pic de $\hat{f}$. Une deuxième composante connexe "nait" à $\alpha = \hat{f}(p_2)$ correspondant à un second pic, mais elle meurt lorsque l'on arrive à la prochaine vallée, c'est-à-dire à $\alpha = \hat{f}(v_1)$, et est fusionnée avec la composante connexe du pic p_1. Ainsi à chaque pic on peut associer une durée de vie, ici les pics p_1 et p_5 ont une longue durée de vie alors que p_2, p_3 et p_4 ont une durée de vie plus courte. Sur la Figure 1 B on représente un dendogramme à la verticale, illustrant les naissances et fusions des composantes connexes de $\{\hat{f} \geq \alpha\}$ lorsque α varie de $+\infty$ à 0. On peut représenter les durées de vie des différents pics dans ce qu'on appelle un diagramme de persistance : un ensemble de points dans le plan, où un point (b, d) correspond à une composante connexe qui est née à l'instant b et morte à l'instant d. Les points se situent sous la diagonale $\{y = x\}$ puisque le "temps" α va à l'envers, de $+\infty$ à 0. On représente en Figure 1 C, le diagrammme de persistance de $\hat{f}$. Lorsque l'on utilise ToMATo, l'homologie persistante est mise en oeuvre et le diagramme permet de voir combien de points ont une durée de vie très longue par rapport aux autres : c'est le nombre de points clairement éloignés de la diagonale. Dans notre

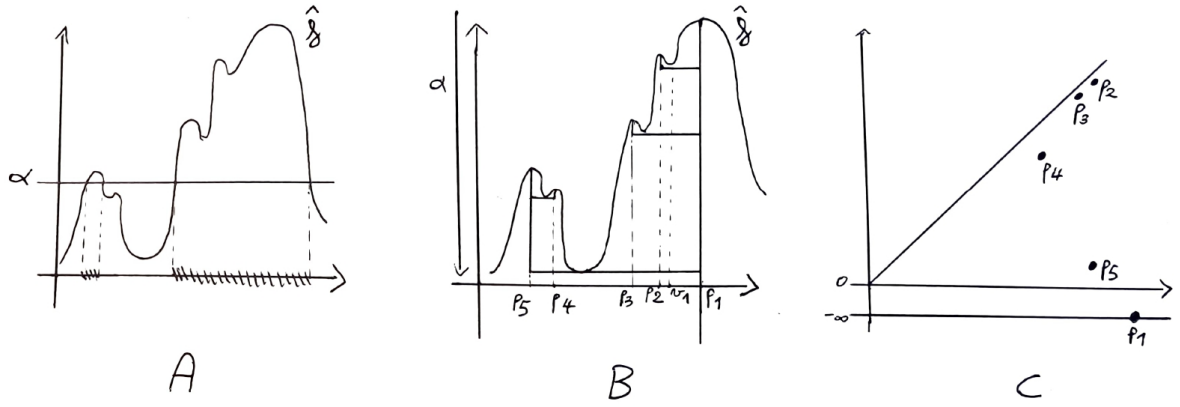


FIGURE 1 – A : le sur-niveau $\{\hat{f} \geq \alpha\}$ pour un certain α. Ce sur-niveau est formé des deux composantes connexes hachurées. B : dendogramme illustrant les naissances et fusions des composantes connexes de $\{\hat{f} \geq \alpha\}$ lorsque α varie de $+\infty$ à 0. Les points correspondant aux pics de $\hat{f}$ sont notés $p_1, \ldots, p_5$. Le point v_1 correspond à une vallée. C : diagramme de persistance de $\hat{f}$. Chaque point correspond à une composante connexe des sur-niveaux de $\hat{f}$, qui correspond aux points tombant sous un pic. On appelle donc également p_i dans le diagramme, le point correspond au pic p_i de $\hat{f}$.

exemple on en compte 2. Avec cette information, on fait tourner ToMATo une seconde fois en précisant que l'on veut 2 clusters (dans notre exemple). L'algorithme effectue à nouveau l'homologie persistante mais cette fois une composante connexe ayant une durée de vie assez grande ne sera jamais fusionnée avec personne et seulement les composantes connexes ayant une courte durée de vie vont être fusionnées, de sorte à obtenir 2 clusters finaux.

Il faut noter que l'estimation de densité n'est pas possible en grande dimension. ToMATo contourne ce problème en utilisant une estimée de la densité seulement en les points des données, puis en construisant un graphe de voisinage. Ensuite, une méthode de *mode-seeking* basé sur les graphes est employée - ToMATo utilise l'algorithme de [Koontz et al., 1976] - et l'homologie persistante est effectuée sur le graphe également. Ainsi ToMATo peut être utilisé même en grande dimension.

2 Applications sur des données simulées

Nous évaluerons la performance et la justesse de ToMATo sur des données simulées de tri des potentiels d'action.

3 Application sur des données réelles

3.1 Description des données

Les données utilisées ici ont été enregistrées dans le premier relais olfactif, le lobe antennaire, du criquet (*Schistocerca americana*). Les conditions d'enregistrement sont détaillées dans l'article [Pouzat et al., 2002]. Les mesures sont réalisées avec une tétrode, ainsi l'activité neuronale est enregistrée à 4 sites différents Dans la figure 2 A, on montre 100 ms de données enregistrées sur les 4 sites d'une tétrode. Les données ont été filtrées entre 300 Hz et 5 kHz avant d'être numérisées avec une fréquence d'échantillonage de 15 kHz. Elles ont ensuite été normalisées en divisant l'amplitude

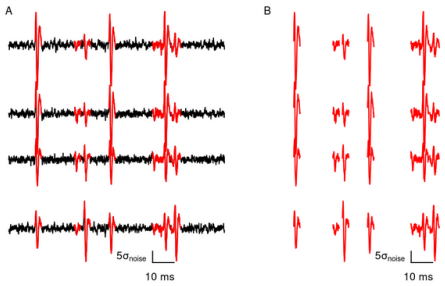


FIGURE 2 – Données d'enregistrement et détection des événements.

du signal sur chaque site par un estimateur robuste, la *median absolute deviation*, de l'écart type du bruit σ_{noise}. Dans la Figure 2 B, après détection des potentiels d'action probables comme extrema locaux suffisamment grands en valeur absolue, des fenêtres sont découpées sur chacun des quatre sites, chaque coupe a 45 points d'échantillonage. Ce groupe de quatre fenêtres détermine un événement défini dans l'espace des événements, ici $\mathbb{R}^{180}$, puisque nous avons $4 \times 45 = 180$ amplitudes par événement.

Une fois que l'on a les événements, la procédure de tri des potentiels d'actions expliquée en détails dans [Pouzat and Detorakis, 2014], se résume aux étapes suivantes :

1. **Nettoyage de l'échantillon** :
   Une superposition d'événements correspond à l'activité de plusieurs neurones qui ont émis un potentiel d'action quasiment en même temps. Ces superpositions vont biaiser l'étape de réduction de dimension, c'est pourquoi il est nécessaire d'enlever préalablement les superpositions les plus évidentes. Cette étape est délicate et n'est pas complètement automatisée, elle doit être réalisée par un expert en tri des potentiels d'action. Après cette étape, il ne reste que les "bons" événements, c'est-à-dire ceux qui ne correspondent pas à des superpositions évidentes.
2. **Réduction de dimension** :
   Puisque les événements vivent dans $\mathbb{R}^{180}$, une étape de réduction de dimension est nécessaire en amont d'un partitionnement. Généralement une analyse en composantes principales est réalisée sur l'ensemble des bons événements.
3. **Visualisation dynamique des données pour trouver le bon nombre de clusters** :
   Le logiciel GGobi est utilisé pour une visualisation dynamique en 3D des (projections sur les composantes principales des) données. On en déduit le bon nombre de clusters. Cette étape nécessite également d'être familier avec ce type de données pour discerner correctement les clusters.
4. **Partitionnement automatique** :
   Une méthode de partitionnement automatique est utilisée, avec le nombre de clusters trouvé précédemment. L'algorithme des K-moyennes est celui qui est mis en oeuvre le plus souvent.

Il peut être laborieux de mettre en oeuvre toutes ces étapes, et l'algorithme ToMATo simplifie considérablement la procédure.

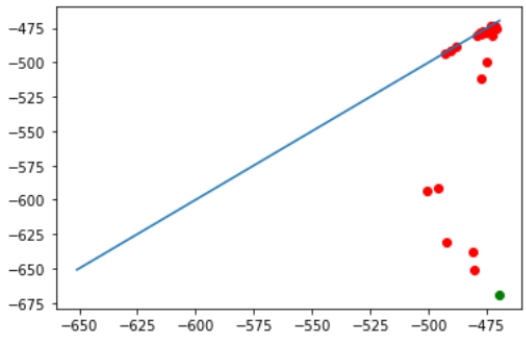


FIGURE 3 – Diagramme de persistance obtenu après avoir appliqué ToMATo sur la collection des événements.

3.2 Application de l'algorithme ToMATo

On montre comment l'algorithme ToMATo permet à lui seul de remplacer les 4 étapes habituelles. Tout d'abord, l'algorithme peut être appliqué directement en grande dimension, ainsi l'étape de réduction de dimension, et donc celle de nettoyage de l'échantillon, ne sont plus nécessaires. En outre, ToMATo fournit un moyen simple de déterminer le nombre de clusters, ainsi les étapes 3 et 4 précédentes sont toutes deux réalisées en utilisant ToMATo.

Nos données sont composées de la collection des événements détectés : nous avons 1950 points dans $\mathbb{R}^{180}$. On utilise l'interface Python de la bibliothèque Gudhi, pour appliquer ToMATo. On fait d'abord tourner l'algorithme une première fois sans spécifier le nombre de clusters, et tous les autres paramètres réglés sur les valeurs par défaut. On obtient le diagramme de persistance affiché en Figure 3. On voit 6 points qui se détachent clairement de la diagonale. On fait alors tourner ToMATo à nouveau en précisant que l'on veut 6 clusters.

ToMATo identifie ainsi 6 neurones ayant généré les événements. On considère qu'il n'y a pas d'information dans la forme de l'événement et on ne retient qu'une information binaire : si le neurone a émis un potentiel d'action ou non. Pour chaque neurone on obtient donc une suite de temps auxquels il a émis un potentiel d'action. Ce résultat est présenté sous la forme de ce qu'on appelle un *rasterplot*, en Figure 4. Chaque ligne correspond à un neurone et sur chaque ligne on a une suite de bâtons, chaque bâton apparaissant à un temps où le neurone a émis un potentiel d'action.

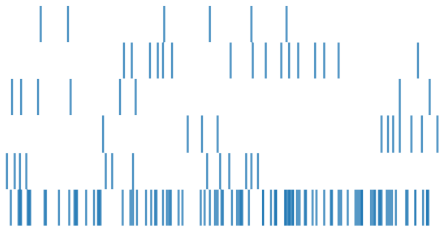


FIGURE 4 – Les 2 premières secondes du *rasterplot* obtenu après le partitionnement.

Bibliographie

Références

[Chazal et al., 2013] Chazal, F., Guibas, L. J., Oudot, S. Y., and Skraba, P. (2013). Persistence-Based Clustering in Riemannian Manifolds. *J. ACM*, 60(6) :1–38.

[Koontz et al., 1976] Koontz, Narendra, and Fukunaga (1976). A Graph-Theoretic Approach to Nonparametric Cluster Analysis. *IEEE Trans. Comput.*, C-25(9) :936–944.

[Pouzat and Detorakis, 2014] Pouzat, C. and Detorakis, G. I. (2014). Spysort : Neuronal spike sorting with python. Proceedings of the 7th European Conference on Python in Science (EuroSciPy 2014), pages 27–34.

[Pouzat et al., 2002] Pouzat, C., Mazor, O., and Laurent, G. (2002). Using noise signature to optimize spike-sorting and to assess neuronal classification quality. *J. Neurosci. Methods*, 122(1) :43–57.

Statistique appliquée à la consommation énergétique

Explainability of electricity demand forecasting models: a Shapley value approach for positive component decomposition

Gaspard Berthelier [1] & Margaux Brégère [2]

[1] *INRIA Sophia-Antipolis and EDF R&D, Lab Paris-Saclay, France* [2] *LPSM, Sorbonne University, Paris, France and EDF R&D, Lab Paris-Saclay, France, margaux.bregere@edf.fr*

Résumé. Prévoir correctement la demande d'électricité est essentiel pour garantir la sureté du système électrique. Les modèles statistiques et d'apprentissage automatique sont largement utilisés mais manquent généralement d'explicabilité. Nous proposons une approche basée sur les valeurs de Shapley pour décomposer les prévisions de la demande d'électricité en composantes positives. Cette décomposition peut être désagrégée, avec une composante pour chaque caractéristique, ou agrégée ; avec, par exemple, deux composantes : l'une pour les effets climatiques (qui représenterait la consommation des usages sensibles aux conditions météorologiques tels que le chauffage et la climatisation) et l'autre pour les effets non climatiques (la consommation des usages électriques non sensibles aux conditions météorologiques). Une telle modélisation pourrait être très utile sur le plan opérationnel pour mieux comprendre l'origine des erreurs de prévision (mauvaises prévisions météorologiques, modélisation non pertinente, etc.), mais aussi pour interpréter l'évolution de la courbe de demande d'électricité, quantifier les éco-actions, les changements de comportement de consommation...

Mots-clés. Explicabilité, valeurs de Shapley, prévision de la demande électrique

Abstract. Accurate electricity demand forecasts are essential to guarantee grid performance and stability. Statistical and machine learning models are widely used but generally lack of explainability. We propose a Shapley value based approach to decompose electricity demand forecasting into positive components. This decomposition may be dis-aggregated, with a component for each feature, or aggregated; with, for example, two parts: one for climatic effects (which would represent the consumption of weather-sensitive uses such as heating and air conditioning) and the other for non-climatic effects (the consumption of non-weather-sensitive uses). Such modelling could be very useful operationally to better understand the origin of forecasting errors (poor weather forecasts, irrelevant modelling, etc.), but also to interpret changes in the electricity demand curve, quantify eco-actions, changes in consumption behaviour, etc.

Keywords. Explainability, Shapley values, Electricity demand forecasting

1 Introduction

Electricity is difficult to store at large scale, except at prohibitive cost, which is why the balance between generation and consumption must be strictly maintained at all times. It is traditionally managed by anticipating the demand as well as the intermittent production (wind, solar), in order to adjust the flexible production (hydro dams, nuclear power plants, coal and gas plants) accordingly. Accurate forecasts of electricity demand and renewable production are therefore essential to guarantee grid performance and stability. State-of-the-art demand forecasting models are complex online machine learning or statistical models that take weather and calendar variables as inputs. Explainability is needed to understand the models, analyse forecasting errors and interpret changes in the demand curve (quantification of eco-actions, penetration of new uses such as electric vehicles, etc.). In practice, this is currently only feasible if the model proposes an intrinsic decomposition according to each variable. That is one of the reasons why Generative Additive Models (GAMs, see among others Pierrot and Goude, 2011) are mostly used, since their additive structure can be used for explainability. Yet, some recent and powerful models such as deep neural networks (Keisler et al., 2024) actually suffer from a lack of interpretability: they are no longer additive are present a huge number of parameters. We propose a Shapley value approach to decompose the forecasts in positive components.

2 Methodology

2.1 Shapley values

Shapley values were introduced as a method to fairly distribute the payoff of a collaborative game among a set of players (Shapley, 1953). They consist in computing the average marginal impact of a player to the game, when considering all the possible subsets (coalitions) of players. Later, they were adapted to machine learning model explainability, by considering features as players and the payoff as the model's output, as seen in Cohen et al. (2007). This is not straightforward. Indeed, let v a game and $J = \{1, \dots, p\}$ a set of players . After the game is played, the players obtain a payoff $v(J)$ which should be distributed fairly by computing the contribution ϕ_j of each player $j \in J$. The unique attribution satisfying fairness properties is (see Shapley, 1953):

$$\phi_j = \sum_{S \subset J \setminus \{j\}} \pi_S (v(S \cup \{j\}) - v(S)) \quad \text{with} \quad \pi_S = \frac{1}{p} \binom{p-1}{|S|}^{-1}.$$

In a machine learning setting, we can consider the local prediction as a game where the players are the feature values $(x_1, \ldots, x_p)$ and the payoff is the prediction $f(x)$. However, how do we define the outcome of the game when a subset of players are absent since f necessarily takes p inputs. Let's note $v_x(S)$ the outcome of the game associated with $f(x)$, when only S is considered present. Various choices for v_x have been proposed in the literature, depending on the interpretation we desire (Rozemberczki et al., 2022, Olsen et al., 2024). In our case, with X_S the random feature vector $(X_j)_{j\in S}$, we consider three versions:

- Baseline: $v_x(S) = f(x_S, z_{\overline{S}})$ where z is a reference point (e.g $z_j = \mathbb{E}[X_j]$)
- Interventional: $v_x(S) = \mathbb{E}[f(x_S, X_{\overline{S}})]$
- Conditional: $v_x(S) = \mathbb{E}[f(X_S, X_{\overline{S}})|X_S = x_S]$.

Both the baseline and interventional method have an issue which is the use of potentially unrealistic data points. Indeed, if features from S and $\overline{S}$ are not independent, a fixed x_S value may be associated with an unlikely $x_{\overline{S}}$ value. For instance, the temperature and calendar variables are highly correlated features. Integrating over the whole scope of temperatures at a given point in time may result in combinations such as (month = January, T = 30°C), which is fairly unrealistic (for now at least). The conditional method appears to be more relevant since it samples $X_{\overline{S}}$ values conditionally to $(X_S = x_S)$. But this requires the knowledge of the conditional distribution $P_{X_{\overline{S}}|X_S=x_S}$ for all coalitions S, which greatly increases the complexity. For the interventional and conditional Shapley values, we approximate the expectations using Monte Carlo methods.

2.2 Post-processing

Aggregating. Once we have obtained trustworthy estimations of the Shapley value of each individual feature, we might like to summarize our results with a reduced number of distinct contributions, for instance a "climate" and "non-climate" component. We partition the set features (e.g $J = C \cup NC$) and introduce the new notation $\tilde{J}$ for the set of reduced coalitions (e.g $\tilde{J} = \{\{C\}, \{NC\}\}$). To compute the Shapley value of $C \in \tilde{J}$, we can simply sum the correspond individual Shapley values:

$$\phi_C(x) = \sum_{j\in C} \phi_j = \sum_{j\in C} \sum_{S\in\mathcal{P}(J\setminus\{j\})} \pi_S\big(v_x(S\cup\{j\}) - v_x(S)\big) \quad \text{with} \quad \pi_S = \frac{1}{p}\binom{p-1}{|S|}^{-1}.$$

This method considers the contribution of a coalition to be the sum of the individual contributions and may not take into account all the joint effects, such as terms of the form

$\pi_S(v_x(S\cup C)-v_x(S))$. Owen (1977) introduced the Coalitional value, which extends Shapley values in the case of coalitional games. These are very useful for specific data structures such as images. To obtain the sole contribution of C, we may use the following simplified formula, inspired by Amoukou et al. (2022), where C acts as a single player playing against all the others:

$$\phi_C(x) = \sum_{S\in\mathcal{P}(J\setminus C)} \pi_S^C\big(v_x(S\cup C) - v_x(S)\big) \quad \text{with} \quad \pi_S^C = \frac{1}{p-|C|+1}\binom{p-|C|}{|S|}^{-1}.$$

The formula can be even more simplified by considering only two players C and NC:

$$\phi_C(x) = \frac{1}{2}\big([v_x(C) - v_x(\emptyset)] + [v_x(J) - v_x(NC)]\big)\,.$$

Heightening and normalizing. Using one of the above-mentioned methods, we have obtained a set of Shapley values. However, they may be negative, which does not make them suitable yet. Indeed, we expect explainable contributions to be positive, since the physical quantity behind is (the electricity demand). Some heightening and normalizing steps allow to get percentage values of the predictions. The whole pipeline is summarized in Figure 1. The XPC (eXplainability through Positive Contributions) `Python` package has been developed to compute the decomposition of forecasts produced by any `R` or `Python` model and codes are open source (`https://github.com/3gaspo/xpc`). It has been used for the experiments below. It is intended to be updated to improve the pipeline and offer more experimental results.

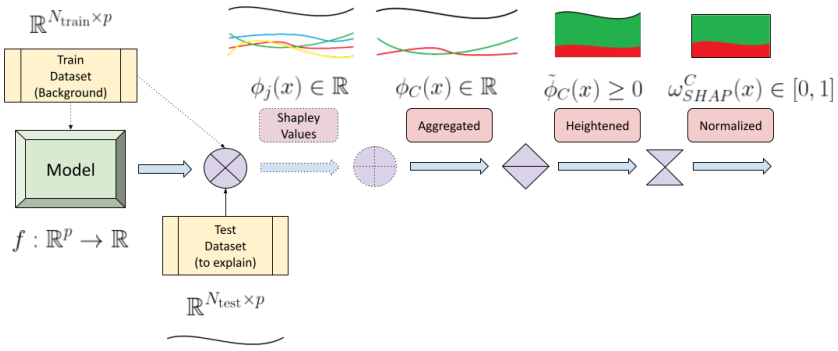


Figure 1: Pipeline to obtain climate and non climate parts of the model predictions.

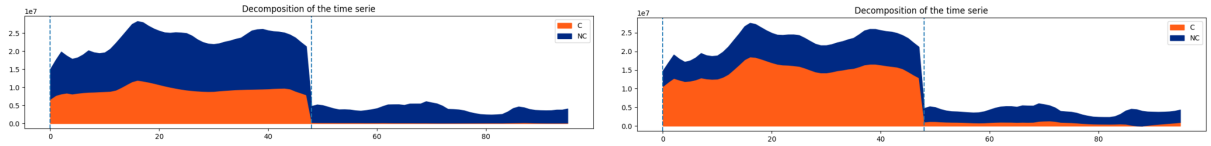


Figure 2: Real decomposition $\phi_C(x)/\phi_{NC}(x)$ of the demand of 5.000 households in Watt given by SMACH data and approximation of the component using the GAM functions.

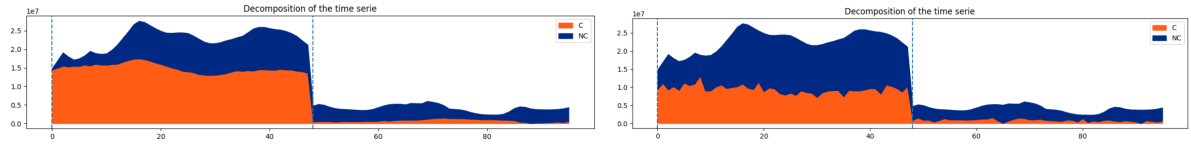


Figure 3: Interventional (left) and conditional (right) Shapley value approximations for a day in winter and a day in summer of the demand of 5.000 households, in Watt.

3 Experiments

3.1 On simulated data: a (rare) case of supervised learning

The dataset was simulated using the SMACH platform (Simulation Multi-Agent de l'Activité humaine et des Consommations dans l'Habitat, Albouys et al., 2019), which simulates the behaviors of numerous households in accordance with French behavioral data collected by the INSEE (Institut National de la Statistique et des Etudes Economiques). This dataset is ideal for explainability: if the forecasting model well-performs, we expect the decomposition in climate (C) and non-climate (NC) components to approximate the demand of weather-sensitive electrical devices (heating and AC) and the one of non-weather-sensitive devices, respectively. We forecast the demand Y using an easily explainable GAM. Indeed, it models the demand as the sum of smooth effects of the weather explanatory variables $(X_j)_{j\in C}$ and the non-weather explanatory variables $(X_j)_{j\in NC}$ and the forecast $\hat{Y}$ can be written:

$$\hat{Y} = \hat{f}(X) = \sum_{j\in C} \hat{f}_j(X_j) + \sum_{j\in NC} \hat{f}_j(X_j)\,,$$

with $\hat{f}_j(\cdot)$ some spline functions (C^2 and piece-wise polynomials) to estimate. The natural decomposition is thus $\hat{\phi}_C(x) = \sum_{j\in C} \hat{f}_j(x_j)$ and $\hat{\phi}_{NC}(x) = \sum_{j\in NC} \hat{f}_j(x_j)$. It is compared to the "real" decomposition $\phi_C(x)/\phi_{NC}(x)$ given by SMACH data (heating and AC / other devices) in Figure 2 for a day in winter and a day in summer. It clearly appears that $\hat{\phi}_C(x)$ is overestimated, this is probably due to the correlation between weather and non-weather variables. We then compute the interventional and conditional Shapley values and plot them in Figure 3 and it is clear that the conditional method does not overestimate the climate part.

On the other hand, the approximation leads to erratic components; which may be reduced by increasing the number of samples in the Monte Carlo pipeline. There is indeed a trade off between precision and inference time.

3.2 On open source data

We then apply our approach on the real open source demand data published by the French Transmission System Operator (`https://www.rte-france.com/eco2mix`) using a large variety of statistical and machine learning models.

References

Albouys, J., Sabouret, N., Haradji, Y., Schumann, M., and Inard, C. (2019). Smach: Multi-agent simulation of human activity in households. In *17th International Conference on Practical Applications of Agents and Multi-Agent Systems (PAAMS 2019)*, pages 227–231.

Amoukou, S. I., Brunel, N. J.-B., and Salaün, T. (2022). The shapley value of coalition of variables provides better explanations.

Cohen, S., Dror, G., and Ruppin, E. (2007). Feature selection via coalitional game theory. *Neural Computation*, 19(7):1939–1961.

Keisler, J., Claudel, S., Cabriel, G., and Brégère, M. (2024). Automated deep learning for load forecasting. *Archive ouverte HAL*.

Olsen, L. H. B., Glad, I. K., Jullum, M., and Aas, K. (2024). A comparative study of methods for estimating model-agnostic shapley value explanations. *Data Mining and Knowledge Discovery*.

Owen, G. (1977). Values of games with a priori unions.

Pierrot, A. and Goude, Y. (2011). Short-term electricity load forecasting with generalized additive models. *Proceedings of ISAP power*, 2011.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., and Sarkar, R. (2022). The shapley value in machine learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5572–5579. International Joint Conferences on Artificial Intelligence Organization.

Shapley, L. S. (1953). *A Value for n-Person Games*. Princeton University Press, Princeton, NJ.

PRÉDICTION DE LA CONSOMMATION ÉNERGÉTIQUE DES TRAINS EN STATIONNEMENT : UNE APPROCHE ÉCONOMÉTRIQUE BASÉE SUR LE MODÈLE PVAR

Nathanaël Kala[1], Myriam Bertrand[2], Frédéric Bertrand[3], Matthieu Leveque[4]

[1] *Université de Technologie de Troyes, France, nathanael_gabriel.kala@utt.fr*
[2] *Ecole des hautes études en santé publique, France, myriam.maumy@ehesp.fr*
[3] *Université de Technologie de Troyes, France, frederic.bertrand@utt.fr*
[4]SNCF Voyageurs, France, *matthieu.leveque@sncf.fr*

Résumé. Cette étude explore la prévision de la consommation énergétique liée au stationnement non conforme des trains dans le contexte spécifique du pôle TER de la SNCF. En s'appuyant sur des données longitudinales provenant de plusieurs sites de dépôt, l'étude utilise le modèle Panel Vector Autoregression (PVAR) pour analyser les dynamiques complexes entre les sites, en tenant compte des variations temporelles et des spécificités intra-sites. Les résultats obtenus permettent d'identifier les périodes à risque de stationnement prolongé, d'optimiser la planification des opérations, et de contribuer à la stratégie de réduction des coûts de la SNCF. Cette étude met également en lumière l'importance d'intégrer des variables supplémentaires et d'examiner l'application de modèles spécifiques à chaque site pour mieux capturer les particularités locales. Les conclusions offrent des perspectives pour l'amélioration continue de la gestion des ressources.

Mots-clés. Modèle économétrique, prédiction, consommation énergétique, stationnement non conforme, PVAR

1 Introduction

Le transport ferroviaire, alternative efficace et peu coûteuse au transport routier, contribue significativement à la réduction de la pollution atmosphérique et de la congestion. Par son activité dense de 15 000 trains par jour, ses 28 000 km de voies ferrées (SNCF Voyageurs, 2024), et son implantation nationale, la SNCF s'affirme comme un acteur essentiel du secteur ferroviaire en France.

Cependant, dans un contexte marqué par la flambée des prix de l'énergie, amplifiée par la faible disponibilité du nucléaire et l'inflation sur le gaz (33, 2024), le gouvernement a appelé les industriels, dont la SNCF, à réduire leur consommation énergétique de 10%. La SNCF, première consommatrice industrielle d'électricité en France avec plus de 9 000 GWhs annuels, doit relever un défi majeur. Si 81 % de cette consommation est dédiée à la traction des trains, l'éco-conduite et l'éco-stationnement sont promus comme solutions clés.

Cette situation pose des enjeux économiques, environnementaux et stratégiques pour la SNCF, affectant potentiellement son efficience opérationnelle et son image. Dans cette étude, nous choisissons de concentrer nos efforts sur les écarts liés aux stationnements non conformes, afin de proposer une stratégie spécifique d'optimisation énergétique adaptée à ce problème. Sachant cela, une question se pose : **comment la SNCF peut-elle utiliser des modèles prédictifs pour anticiper et réduire les consommations d'énergie causées par le stationnement non conforme ?**

2 Méthodologie

2.1 Description des données

Pour cette étude nous utilisons un ensemble de données mensuelles couvrant les sites de dépôts du matériel roulant (gares et technicentres) des lignes couvertes par la DLPGE (Direction de Ligne Paris Grand Est). La période étudiée s'étend de janvier 2021 à juin 2024, offrant ainsi une longue perspective chronologique.

Nous avons sélectionné deux indicateurs clés pour cette étude : les durées cumulées de stationnement en énergie électrique et en énergie thermique. Ces indicateurs sont complétés par les données de températures moyennes journalières, recueillies via l'API Open Data de Météo France. Les données relatives aux durées de stationnement proviennent du projet PlaneTER.

Les consommations d'énergie associées sont calculées en fonction du mode d'énergie utilisé :

- Cumul d'exploitation en énergie électrique (CE_EE) :

$$CE_EE = CONSO_kWh_STA \times \left(\frac{DSE}{60} \right) (1)$$

  - *CONSO_kWh_STA* représente la consommation moyenne en kWh en maintien de service, estimée à 70 kWh ;
  - *DSE* est la variable représentant la durée cumulée de stationnement en énergie électrique, exprimée en minutes.
- Cumul d'exploitation en énergie thermique (CE_ET) :

$$CE_ET = CONSO_LH_PWP_STA \times \left(\frac{DST}{60} \right) (2)$$

  - *CONSO_LH_PWP_STA* représente la consommation moyenne en litres par heure par moteur diesel (un seul étant en service lorsque l'engin moteur est en mode économie d'énergie) ;
  - *DST* est la variable représentant la durée cumulée de stationnement en énergie thermique, exprimée en minutes.

2.2 Spécification du modèle

Le modèle Panel Vector Autoregression (PVAR) adapté à nos variables "durée de stationnement en électrique" (stationnement_elec) et "durée de stationnement en thermique" (stationnement_ther) pour les variables dépendantes, ainsi qu'à la moyenne des températures par mois (TNTXM) pour la variable exogène, peut être spécifié comme suit :

$$stãtio_elec_{i,t} = \sum_{p=1}^{P} \phi_{11,p} \cdot stãtio_elec_{i,t-p} + \sum_{p=1}^{P} \phi_{12,p} \cdot stãtio_ther_{i,t-p} + \gamma_1 \cdot TNT\tilde{X}M_{i,t} + \tilde{u}_{1,i,t} \quad (3)$$

$$stãtio_ther_{i,t} = \sum_{p=1}^{P} \phi_{11,p} \cdot stãtio_elec_{i,t-p} + \sum_{p=1}^{P} \phi_{12,p} \cdot stãtio_ther_{i,t-p} + \gamma_2 \cdot TNT\tilde{X}M_{i,t} + \tilde{u}_{1,i,t} \quad (4)$$

où :

- $stãtio_elec_{i,t}$, $stãtio_ther_{i,t}$ et $TNT\tilde{X}M_{i,t}$ sont les variables centrées par les moyennes.
- $\phi_{11,p}$, $\phi_{12,p}$, $\phi_{21,p}$, et $\phi_{22,p}$ sont les coefficients autorégressifs, capturant les relations dynamiques entre les variables après transformation.
- γ_1 et γ_2 sont les coefficients qui capturent l'impact direct de la température moyenne $TNT\tilde{X}M_{i,t}$ sur les durées de stationnement en mode électrique et thermique respectivement.

- $\tilde{u}_{1,i,t}$ et $\tilde{u}_{2,i,t}$ sont les termes d'erreur idiosyncratiques après transformation.

2.3 Validation du modèle : test de racines unitaires et analyse de cointégration

Les résultats des trois tests de stationnarité IPS (Im et al., 2003), LLC (Levin et al., 2002) et MW (Maddala & Wu, 1999) indiquent clairement que les séries de données relatives aux variables de durée de stationnement sont stationnaires en niveau, sans nécessiter de transformation en différences premières. De meme, les résultats des trois tests de cointégration Kao (1999), Pedroni (2004) et Westerlund (2005) rejettent de manière significative l'hypothèse nulle d'absence de cointégration.

En ce qui concerne la sélection du nombre optimal de retards, celui de notre modèle PVAR(p) est basée sur des critères de sélection, tels que le critère d'information d'Akaike (AIC) et le critère d'information bayésien (BIC). Sur la base de ces critères, nous nous rendons compte de la diminution de ces critères au fur et à mesure que le lag sélectionné est grand, du coup pour éviter un surparamétrage de notre modèle, nous allons nous limiter à p=9.

2.4 Démarche de prédiction

La prédiction avec un modèle PVAR consiste à utiliser les dynamiques temporelles capturées par le modèle pour anticiper les valeurs futures des variables d'intérêt. Dans le cadre de cette étude, les prédictions ont été réalisées sur les durées cumulées de stationnement en énergie électrique et en énergie thermique pour chaque site de dépôt, en s'appuyant sur les estimations des coefficients obtenus à partir des données historiques.

La démarche de prédiction a consisté à appliquer les coefficients estimés aux données des périodes précédentes pour anticiper les valeurs des périodes futures. Les prévisions out-of-sample ont été réalisées en utilisant les valeurs laggées des variables explicatives (durées de stationnement en énergie électrique et thermique) et de la variable exogène (température moyenne, TNTXM) pour projeter les durées cumulées de stationnement pour les mois suivant la période d'étude. Les prédictions ont été calculées pour chaque site de dépôt en utilisant la spécification des équations (3) & (4). Les prédictions ont été ensuite recentrées en ajoutant la moyenne de chaque variable sur l'ensemble de la période d'étude, pour obtenir les niveaux absolus des durées de stationnement prédites.

3. Résultats empiriques & implications pratiques

Initialement, nous avons envisagé d'utiliser le pourcentage d'erreur absolu moyen (MAPE), l'une des mesures les plus couramment utilisées pour évaluer la précision des modèles. Cependant, la présence significative de valeurs nulles dans nos observations, notamment pour la variable de stationnement électrique, a rendu le MAPE inapproprié, conduisant à des valeurs infinies dans nos calculs.

Pour surmonter cette limitation, nous nous sommes tournés vers une mesure alternative proposée par (Kim & Kim, 2016), le Mean Arctangent Absolute Percentage Error (MAAPE). Le MAAPE, bien que moins connu que le MAPE, présente l'avantage de gérer efficacement les valeurs nulles, offrant ainsi une évaluation plus robuste et fiable de la précision du modèle dans notre contexte.

Les résultats obtenus à partir du MAAPE pour nos prédictions (Tableau 5) révèlent des performances variées selon les sites et les types de stationnement. Pour le stationnement thermique, le MAAPE global est de **53.96%**, indiquant une précision modérée à faible.

Toutefois, des disparités importantes sont observées entre les sites, avec des valeurs oscillantes entre **22.35%** à Troyes, où le modèle semble bien capturer les dynamiques, et **85.14%** à Châlons-en-Champagne, où la précision est beaucoup plus faible.

Variables	Sites	MAAPE
Stationnement thermique	**Total**	**53,96**
	AG Châlons-en-Champagne	48,31
	Châlons-en-Champagne	85,14
	Culmont-Chalindrey	71,38
	Saint-Dizier	76,13
	Technicentre de Chalindrey	23,00
	Technicentre de Mulhouse	73,81
	Technicentre de Pantin	31,56
	Troyes	22,35
Stationnement électrique	**Total**	**84,72**
	AG Châlons-en-Champagne	97,47
	Châlons-en-Champagne	93,19
	Culmont-Chalindrey	110,33
	Saint-Dizier	104,47
	Technicentre de Chalindrey	94,91
	Technicentre de Mulhouse	60,07
	Technicentre de Pantin	15,42
	Troyes	101,95

Tableau 1. Résultats de précisions : Mean Arctangent Absolute Percentage Error (MAAPE)

Concernant le stationnement électrique, le MAAPE global est plus élevé, à **84.72%**, suggérant des prédictions nettement moins précises dans ce contexte. Là encore, les résultats varient considérablement d'un site à l'autre, avec un MAAPE particulièrement faible de **15.42%** pour le Technicentre de Pantin, tandis que d'autres sites, comme Culmont-Chalindrey avec un MAAPE de **110.33%**, montrent que le modèle a des difficultés à prédire avec précision les valeurs observées.

Ces résultats peuvent en partie s'expliquer par la prédominance de l'une des variables (électrique ou thermique) sur chaque site, mis à part le Technicentre de Pantin. Par exemple, à Troyes, les valeurs de stationnement thermique sont consistantes et montrent une bonne variance, tandis que les valeurs électriques sont quasiment nulles. Cette prédominance crée une situation où le modèle, en raison de l'approche de centrage utilisée, se trouve limité dans sa capacité à généraliser efficacement sur les deux types de stationnement, car les dynamiques d'une variable ne sont pas suffisamment présentes pour être modélisées avec précision.

Bien que les valeurs du MAAPE fournissent un aperçu global de la précision du modèle, une analyse graphique des prédictions permet d'apporter des nuances importantes à cette évaluation. En examinant les courbes de prédictions *in-sample* (prédictions réalisées sur l'ensemble de données d'entraînement) et *out-of-sample* (prédictions futures), certaines tendances et comportements spécifiques émergent (Figures 1 et 2).

Pour le **stationnement électrique**, le Technicentre de Pantin se distingue par une bonne correspondance entre les valeurs observées et prédites. Ce résultat est encourageant et suggère que le modèle peut être fiable dans des contextes où les dynamiques sont plus stables ou moins sujettes à des perturbations. Cependant, d'autres sites comme Saint-Dizier ou Châlons-en-Champagne montrent des écarts importants entre les valeurs observées et les prédictions, ce qui met en évidence les limitations du modèle dans des environnements plus complexes ou instables. Ces résultats soulignent la nécessité d'approfondir l'analyse pour identifier les facteurs spécifiques qui peuvent expliquer ces divergences et pour envisager des améliorations du modèle.

Figure 1. Prédictions sur les durées de stationnement non conforme en électrique

Figure 2. Prédictions sur les durées de stationnement non conforme en thermique

L'analyse graphique permet également de constater que, malgré les faiblesses du modèle sur certains sites, il réussit à capturer les tendances globales pour d'autres. Cette observation souligne l'importance de ne pas se fier uniquement aux mesures globales de précision comme le MAAPE, mais d'intégrer une analyse visuelle pour mieux comprendre les performances du modèle dans différents contextes.

4. Conclusion

Cette étude visait à démontrer l'importance de la prévision des charges opérationnelles liées au stationnement non conforme des trains, en se focalisant sur le contexte particulier de la SNCF,

et plus précisément sur le pôle TER. L'utilisation du modèle économétrique Panel Vector Autoregression (PVAR) a permis de capturer, du moins en partie, les dynamiques complexes propres à chaque site de stationnement, en prenant en compte les variations temporelles et les spécificités intra-site.

L'entreprise pourrait tirer parti de cette étude pour anticiper les périodes à risque de stationnement prolongé, optimiser la planification des opérations, et ainsi contribuer à la stratégie globale de réduction des coûts de la SNCF sur les sites visés De plus, ce travail apportant une contribution à la littérature existante, les recherches futures devraient viser à enrichir les données en intégrant des variables supplémentaires pour mieux appréhender les particularités de chaque site, et à étudier l'application de modèles spécifiques lorsque l'hétérogénéité entre les sites est trop marquée.

Bibliographie

Canova, F., & Ciccarelli, M. (2013). Panel Vector Autoregressive Models: A Survey☆ The views expressed in this article are those of the authors and do not necessarily reflect those of the ECB or the Eurosystem. In *VAR models in macroeconomics–new developments and applications: Essays in honor of Christopher A. Sims* (pp. 205-246). Emerald Group Publishing Limited.Im, K. S., Pesaran, M. H., & Shin, Y. (2003). Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, *115*(1), 53-74.

Im, K. S., Pesaran, M. H., & Shin, Y. (2003). Testing for unit roots in heterogeneous panels. *Journal of econometrics*, *115*(1), 53-74.
Kao, C. (1999). Spurious regression and residual-based tests for cointegration in panel data. *Journal of Econometrics*, *90*(1), 1-44.

Maddala, G. S., & Wu, S. (1999). A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test. *Oxford Bulletin of Economics and Statistics*, *61*(S1), 631-652.

Pedroni, P. (2004). Panel cointegration: asymptotic and finite sample properties of pooled time series tests with an application to the PPP hypothesis. *Econometric theory*, *20*(3), 597-625.

SNCF Voyageurs. (2024, février 28). Groupe SNCF. https://www.groupe-sncf.com/fr/groupe/portrait-entreprise/groupe-societes/sncf-voyageurs

Stock, J. H., & Watson, M. W. (2001). Vector Autoregressions. *Journal of Economic Perspectives*, *15*(4), 101-115.

Westerlund, J. (2005). A Panel CUSUM Test of the Null of Cointegration. *Oxford Bulletin of Economics and Statistics*, *67*(2), 231-262.

MODÉLISATION ET SIMULATION DES DÉFAILLANCES DANS LE RÉSEAU DE DISTRIBUTION ÉLECTRIQUE DE LA CIE : CAS DU POSTE SOURCE DE LA RIVIERA

Marcellin Brou[4], Efoevi Koudou[5], Arnold N'goran[2], Gaetan Otogo[1], Minayégnan Touré[3], Anne-Françoise Yao[4,7],

[1] *UMRI, EDP-INPHB, Côte d'Ivoire, otogo.ndoutoume24@inphb.ci*
[2] *CIE SDRO, Côte d'Ivoire, arnold.ngoran@cie.ci*
[3] *CIE DCPIQ, Côte d'Ivoire, mitoure@cie.ci*
[4] *Université Clermont Auvergne, CNRS, LMBP, CNRS, Aubière Cedex, France, anne.yao@uca.fr*
[5] *INPHB, Côte d'Ivoire, marcellin.brou@inphb.ci*
[6] *Université de Lorraine, CNRS, IECL, F-54000 Nancy, France, efoevi.koudou@univ-lorraine.fr*
[7] *Centre de Mathématiques Appliquées, Ecole polytechnique, Paris, France, anne.yao@uca.fr*

Résumé Ce travail explore la modélisation et la simulation des défaillances dans le réseau de distribution électrique à partir des données fournies par la Compagnie Ivoirienne d'Electricité (CIE) sur le réseau électrique. L'étude vise à améliorer la résilience et la fiabilité du réseau électrique en identifiant et en analysant les impacts des défaillances sur la performance globale du réseau. La méthodologie mise en place repose sur la création d'un modèle détaillé du réseau électrique et l'analyse de l'écoulement de puissance via différents scénarios de simulation tels que : les ruptures de lignes ou les pannes d'équipements critiques. Les résultats obtenus mettent en évidence les points faibles du réseau de distribution électrique et permettent de tester des solutions correctives et préventives. Ces simulations, combinées à des approches analytiques avancées, facilitent ainsi le suivi de l'évolution du réseau de distribution électrique et la planification des éventuelles interventions au vue de son amélioration. Cette approche s'inscrit dans le cadre d'une dynamique de transformation du réseau électrique en système intelligents, capables d'anticiper et de gérer efficacement les défaillances. Ce qui souligne l'importance de la modélisation et des simulations numériques dans l'évolution des systèmes électriques modernes.

Mots-clés. Réseau électrique, Défaillance électrique, Modélisation, Pandapower.

Abstract. This study focuses on modeling and simulating failures in the electrical distribution network using data from the Ivorian Electricity Company (CIE). It aims to enhance the network's resilience and reliability by analyzing failure impacts on overall performance. The methodology involves creating a detailed network model and simulating scenarios like line outages or critical equipment failures. Results reveal network weaknesses and enable virtual testing of corrective and preventive measures. Combined with advanced analytics, these simulations support monitoring, planning interventions, and transforming the grid into intelligent systems capable of anticipating and managing failures, highlighting the role of modeling in modern electrical systems.

Keywords. Electrical network, Electrical failure, Modeling, Pandapower.

1 Introduction

Le réseau de distribution électrique constitue une infrastructure essentielle pour garantir un approvisionnement énergétique fiable et durable. Cependant, ces réseaux sont fréquemment confrontés à des défaillances causées par divers facteurs, tels que les pannes d'équipement, les aléas climatiques ou des erreurs humaines, compromettant ainsi leur performance et leur résilience. Dans ce contexte, la Compagnie Ivoirienne d'Électricité (CIE) fournit un cadre propice à l'exploration des mécanismes de défaillance et des solutions innovantes pour leur gestion.

Comment améliorer la résilience et la fiabilité du réseau de distribution électrique en identifiant et en analysant les défaillances tout en testant des solutions correctives et préventives de manière efficace ?

Ce travail vise à renforcer la résilience et la fiabilité du réseau de distribution électrique à partir de données fournies par la CIE, en utilisant des outils de modélisation et de simulation pour comprendre et anticiper les impacts des défaillances sur la performance globale du réseau.

De façon spécifique, cet article vise à :

1. Développer un modèle détaillé du réseau de distribution électrique pour analyser son comportement en situation normale et lors de défaillances.
2. Simuler différents scénarios de pannes, notamment les coupures de lignes et les défaillances critiques d'équipement, afin d'évaluer leurs impacts sur le réseau.
3. Identifier les points faibles du réseau en vue de proposer des solutions correctives et préventives.
4. Contribuer à la transformation du réseau électrique en un système intelligent capable d'anticiper et de gérer efficacement les défaillances.
5. Améliorer la prise de décision en fournissant à la CIE, un outil d'aide à la décision basé sur des analyses fiables et des simulations réalistes.

Ce travail s'inscrit dans une dynamique de modernisation des systèmes électriques, où la modélisation et les simulations numériques jouent un rôle clé pour anticiper les défis de demain tout en assurant une gestion proactive et optimisée du réseau.

2 Méthodologie

2.1 Réseau de distribution électrique

Le réseau électrique est généralement défini comme étant l'ensemble des appareillages mis en œuvre pour faire transiter l'énergie électrique des centres de productions vers les centres de consommation. Il englobe la totalité des installations, notamment les lignes, les transformateurs et les appareils avec leurs moyens de contrôle et de sécurité nécessaires au transport et à la distribution de l'énergie électrique.

2.2 Modélisation du réseau de distribution électrique

La modélisation est une étape nécessaire pour l'analyse du comportement du réseau électrique. Ainsi pour étudier notre réseau électrique complexe, des modèles ou schémas équivalents de ces principaux composants (transformateurs, lignes, nœuds ; charges) sont établis.

Les Transformateurs

Le transformateur est une machine statique qui a pour rôle de transiter une énergie électrique d'un réseau à un autre en modifiant ses caractéristiques : intensité et tension.

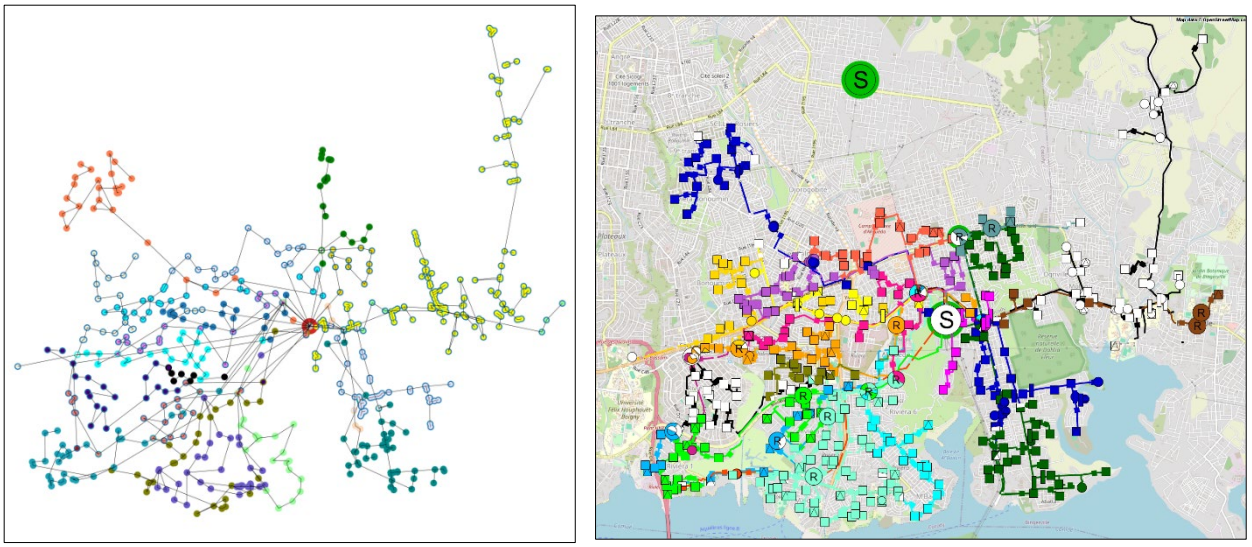
Les Lignes

L'énergie électrique est transitée des centres de productions vers les consommateurs à travers les lignes électriques. Les lignes souterraines sont plus utilisées dans le milieu urbain et les lignes aériennes pour les périphéries et le réseau de transport.

Les Charges

Le fonctionnement stable d'un réseau électrique est assuré par la capacité d'alimenter des charges de manière continue. Les modèles de charge électrique sont généralement divisés en model statique et dynamique. La modélisation dynamique des charges est relativement compliquée car la puissance consommée par la charge est en fonction de la tension et du temps.

Figure 1 : Réseau de distribution électrique du poste source de la riviera.



(a)Visualisation dans PANDAPOWER (b) Visualisation dans ERABLE

2.3 Équations de l'écoulement de puissance

L'analyse de l'écoulement de puissance repose sur des équations mathématiques qui décrivent le comportement électrique d'un réseau de distribution. Ces équations s'appuient sur la modélisation du réseau à l'aide de la matrice d'admittance nodale et sont résolues à l'aide de méthodes numériques adaptées.

Formulation de la matrice d'admittance

La matrice d'admittance nodale ***Y*** représente la relation entre les tensions et les courants des nœuds d'un réseau électrique.

Définition :

$$\boldsymbol{I}_i = \sum_{j=1}^{n} \boldsymbol{Y}_{ij}\boldsymbol{V}_j \ . \tag{1}$$

Où :

- $\boldsymbol{I}_i$: $Courant\ injecté\ au\ noeud\ i.$
- $\boldsymbol{V}_j$: $Tension\ au\ noeud\ j.$
- $\boldsymbol{Y}_{ij}$: $Elément\ (i,j)\ de\ la\ matrice\ d'admittance.$

Construction :

La matrice $\boldsymbol{Y}$ est construite à partir des admittances de ligne et des admittances shunt.

- $\boldsymbol{Y}_{ii} = \sum_{j\neq i} y_{ij} + \ y_{shunt,i}$ $(éléments\ diagonaux)$.
- $\boldsymbol{Y}_{ij} = -y_{ij}\ (i \neq j), o\grave{u}\ y_{ij}\ \ est\ l'admittancede\ la\ ligne\ entre\ i\ et\ j.$

La matrice $\boldsymbol{Y}$ est fondamentale pour exprimer les relations entre tensions, courants, et puissances dans le réseau.

Formulation des équations de l'écoulement de puissance

Les équations d'écoulement de puissance décrivent les puissances actives ***(P)*** et réactives ***(Q)*** injectées au niveau des nœuds du réseau. Ces puissances sont exprimées comme suit :

- **Puissance Active au nœud *i***

$$\boldsymbol{P}_i = \boldsymbol{V}_i \sum_{j=1}^{n} \boldsymbol{V}_j(\boldsymbol{G}_{ij}\cos\theta_{ij} + \boldsymbol{B}_{ij}\sin\theta_{ij}) \ . \tag{2}$$

- **Puissance Réactive au nœud *i***

$$\boldsymbol{Q}_i = \boldsymbol{V}_i \sum_{j=1}^{n} \boldsymbol{V}_j(\boldsymbol{G}_{ij}\sin\theta_{ij} - \boldsymbol{B}_{ij}\cos\theta_{ij}) \tag{3}$$

Où :

- $\boldsymbol{V}_i, \boldsymbol{V}_j$: $Magnitude\ des\ tensions\ aux\ noeuds\ i\ et\ j.$
- $\boldsymbol{G}_{ij}, \boldsymbol{B}_{ij}$: $Parties\ réelle\ et\ imaginaire\ de\ la\ matrice\ d'admittance\ nodale.$
- $\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_i - \ \boldsymbol{\theta}_j$: $Différence\ d'angle\ entre\ les\ tensions.$

Ces équations sont non linéaires et nécessitent des méthodes numériques pour être résolues.

La méthode de Newton-Raphson

La méthode de Newton-Raphson est la technique la plus couramment utilisée pour l'analyse de l'écoulement de puissance en raison de sa rapidité et de sa précision.

Les équations non linéaires $\boldsymbol{f(x) = 0}$ sont linéarisées en utilisant une expansion en série de Taylor :

$$\boldsymbol{f}(x_k) + \boldsymbol{J}(x_k)\Delta x = 0 \tag{4}$$

Où :

- $\boldsymbol{x_k}$: $est\ le\ vecteur\ des\ variables\ (tensions\ et\ angles)$à $l'itération\ de\ k.$
- $\boldsymbol{J(x_k)}$: $C'est\ la\ Jacobienne, qui\ contient\ les\ dérivées\ partielles\ des\ équations\ par\ rapport\ aux\ variables.$
- $\Delta \boldsymbol{x} = \boldsymbol{x_{k+1}} - \boldsymbol{x_k}$: $C'est\ le\ vecteur\ de\ correction$

3 PandaPower

PandaPower est un outil d'analyse de systèmes d'alimentation basé sur Python et sous licence BSD destiné à l'automatisation de l'analyse statique et quasi-statique et à l'optimisation de systèmes d'alimentation équilibré. Il existe de nombreuses méthodes itératives utilisées pour résoudre l'ensemble des équations non linéaires trouvées dans les algorithmes de flux de puissance. Les 5 qui sont spécifiés dans PandaPower sont la méthode Newton-Raphson (nr), Newton-Raphson avec multiplicateur Iwamoto, balayage avant/arrière, Gauss-Seidel et découplé rapidement. La méthode la plus populaire utilisée (et en tant que telle la méthode par défaut utilisée dans PandaPower) est la méthode Newton-Raphson.

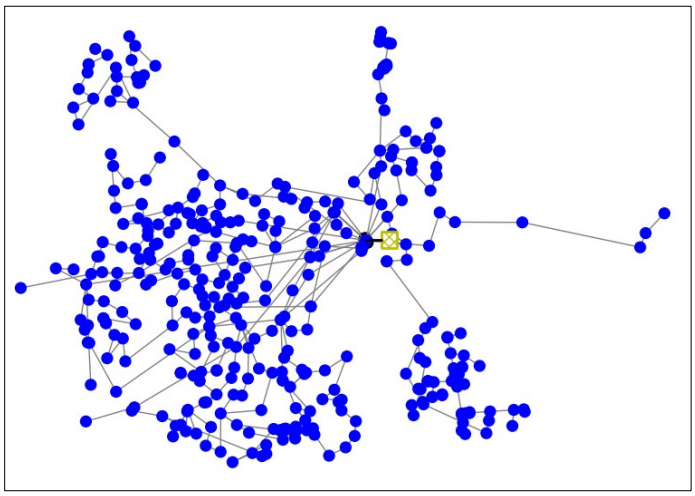


Figure 2 : Modélisation du réseau dans pandapower

4 Simulations et Résultats

Les deux images montrent les résultats d'une simulation de flux de puissance basée sur la méthode de **Newton-Raphson**, une analyse courante pour comprendre le comportement d'un réseau de distribution électrique de la CIE. La première image met en évidence la tension aux nœuds et le chargement des lignes, tandis que la deuxième image se concentre sur le chargement des lignes.

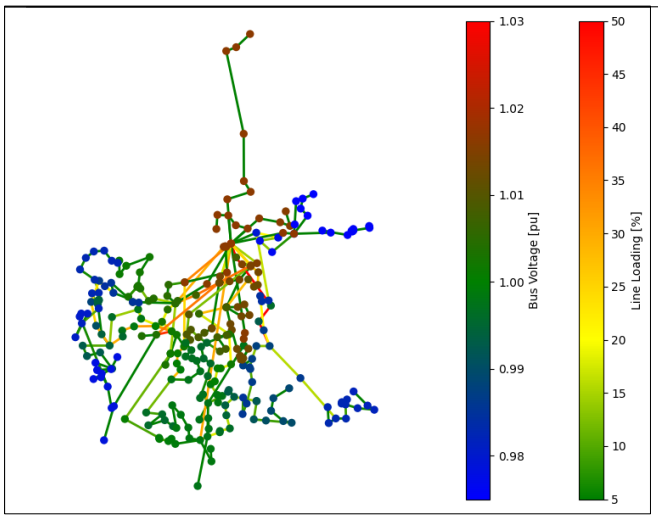


Figure 3 : Tension aux nœuds et charge de ligne

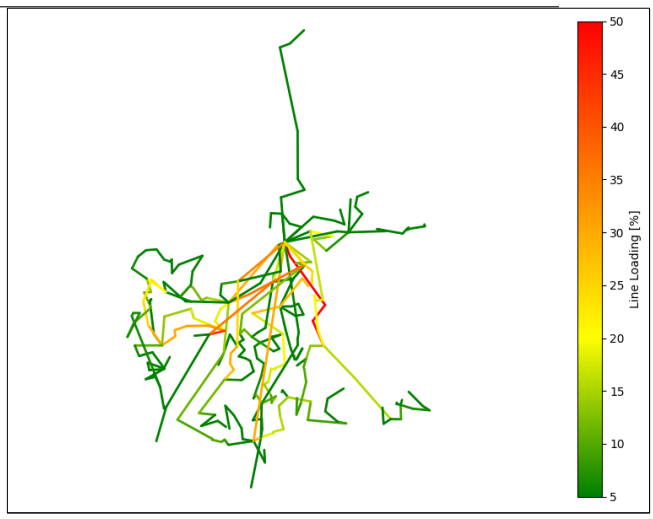


Figure 4 : Charge de ligne.

Tableau 1 : Analyse globale du diagnostic du réseau de distribution.

Critère	**Résultats observés**	**Risques et Implications**
Taux de charge max	> 45 %	Risque de congestion et échauffement
Profil de tension	[0.97 pu – 1.03 pu], dispersé	Problèmes de qualité de tension localisée
Zones critiques	Centre du graphe (nœud d'interconnexion)	Nécessité de renforcer ou décharger

La majorité des lignes conservent un taux de charge modéré à faible (< 25 %). Plusieurs lignes présentent une charge > 35 %, dont au moins deux segments en rouge > 45 %, ce qui montre une mise sous tension importante suite à la redondance perdue. Le centre du graphe est fortement sollicité, typique d'un nœud de transit forcé après perte de chemins alternatifs.

Les zones rouges indiquent des **goulots d'étranglement** en formation : bien qu'encore en dessous de 100 %, elles sont critiques car proches de la **surcharge en régime dynamique** (surtout en cas de montée de charge). Ce comportement est cohérent avec un **réseau peu maillé** ou **avec peu d'interrupteurs ouverts stratégiquement positionnés**, limitant les chemins de secours efficaces.

Les tensions restent **globalement dans les bornes** [0.97 – 1.03 pu], mais :

- Plusieurs **nœuds bleus (0.97 pu)** indiquent des **chutes de tension significatives**.
- Des nœuds **rouge (> 1.02 pu)** indiquent une **surtension probable liée à une injection locale sans absorption suffisante**.

La superposition montre que **les lignes les plus chargées** (**rouges**) sont **proches des zones à basse tension**, ce qui **renforce l'effet Joule** et aggrave la dégradation du profil de tension.

5 Analyse statistique des résultats

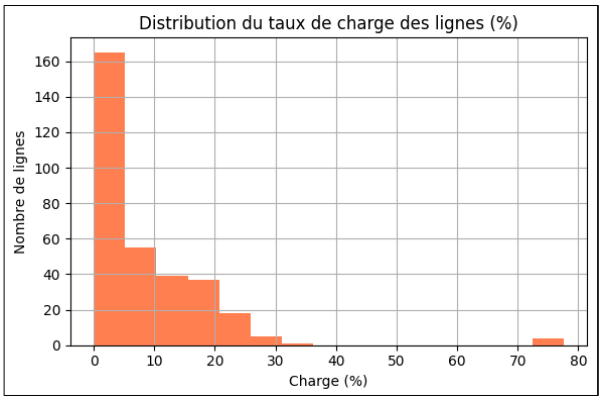


Figure 5 : Distributions de charge des lignes

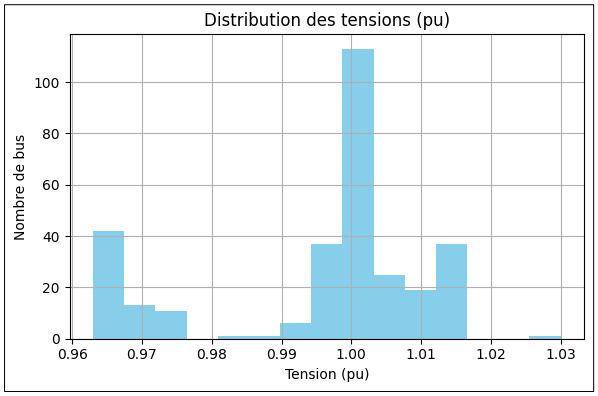


Figure 6 : Distribution des tensions

En complément de l'analyse visuelle, une distribution des charges de lignes a été extraite afin de caractériser la probabilité d'apparition de contraintes thermiques. L'estimation de densité indique que la majorité des lignes sont exploitées à moins de **10 %** de leur capacité, ce qui reflète un fort potentiel de rationalisation des actifs. Toutefois, un petit groupe de lignes dépasse les **70 %** de charge, dont environ **1 à 2 %** atteignent des niveaux proches de **80 %,** pouvant signaler une criticité locale en exploitation normale ou une vulnérabilité accrue en régime dégradé (critère N-1).

Parallèlement, l'analyse de la variabilité des tensions aux nœuds révèle une dispersion relativement faible, avec un écart-type estimé inférieur à **0.015 pu**, traduisant une bonne régulation globale. Néanmoins, la présence de valeurs de tension inférieures à **0.97 pu** et supérieures à **1.01 pu** suggère des comportements divergents selon les zones du réseau, notamment sous l'effet de charges déséquilibrées.

6 Perspectives d'exploitation statistique

Ce type d'analyse statistique, couplé à des scénarios de simulation, constitue une base pertinente pour calibrer des modèles de prévision de contraintes ou de défaillances. Des approches par apprentissage supervisé pourraient être envisagées pour affiner la gestion proactive des risques réseau.

Il est envisageable de construire des modèles statistiques de fiabilité, tels que :

- L'estimation de la probabilité de défaillance conditionnelle sur la base de la topologie et de la charge moyenne.

- L'application d'algorithmes d'apprentissage automatique (arbres de décision, réseaux bayésiens) pour prédire les zones à risque élevé de coupure.

Ce type d'approche statistique, en complément des méthodes déterministes, pourrait enrichir la prise de décision stratégique de la CIE dans une logique proactive.

7 Conclusion

En conclusion, le réseau de distribution électrique est crucial pour assurer un approvisionnement énergétique fiable et durable, mais il est souvent sujet à des défaillances variées. La Compagnie Ivoirienne d'Électricité (CIE) joue un rôle clé en fournissant un cadre pour explorer et gérer ces défaillances. Ce travail vise à améliorer la résilience et la fiabilité du réseau en utilisant des outils de modélisation et de simulation pour anticiper les impacts des défaillances et proposer des solutions correctives et préventives. En développant un modèle détaillé du réseau, en simulant divers scénarios de pannes, et en identifiant les points faibles, ce projet contribue à transformer le réseau en un système intelligent capable de gérer efficacement les défaillances et d'améliorer la prise de décision grâce à des analyses fiables et des simulations réalistes.

Remerciement

Ce travail est réalisé dans le cadre la Chaire Internationale Data Science Institute programme de l'Ecole polytechnique Paris, (France) et l'Institut National Polytechnique Félix Houphouët-Boigny (INPHB) financé par la Compagnie Ivoirienne d'Electricité (CIE)

Bibliographie

Colelough, B. C. 15 Mar 2024. *Australian Energy Market Operator National Electricity Market Network Optimal Power Flow Modelling*.

Compagnie. Ivoirienne. Electricité (CIE). Mars 2016. *Choix techniques et doctrine de la planification des réseaux électriques de distribution en Côte d'Ivoire*.

C.I.E. Juin 2024. *Consultation des données du Système d'Information Géographique des réseaux électriques de distribution en Côte d'Ivoire.*

Oussama, E. Y. 05 Juin 2018. *Étude d'amélioration du réseau HTA issu d'un transformateur T1 (Poste source DAR BOUAZZA départ D10) par la compensation de l'énergie réactive*.

Van der Plas, F. July 2019. *Power grid failures*.

Rishabh Verma1, Vaskar Sarkar1. 11 Octobre 2018. *Active distribution network load flow analysis through non-repetitive FBS iterations with integrated DG and transformer modelling*

L. Thurner et al. 2018. *Pandapower—An Open-Source Python Tool for Convenient Modeling, Analysis, and Optimization of Electric Power Systems IEEE Transactions on Power Systems*, vol. 33 Available at: https://ieeexplore.ieee.org/document/8344496 doi:10.1109/TPWRS.

2018.2829021.

W H Kersting, CRC Press 2002 "Distribution system modeling and analysis, Book,

Jean-Philippe Grivet, Magali Ribot, EDP Sciences, 2013, Méthodes Numériques Appliquées pour le scientifique et l'ingénieur.

DRAIDI Abdellah ; 11 juillet 2016. Répartition optimale des puissances utilisant les techniques de l'intelligence artificielle » Thèse de Doctorat En Électrotechnique ; Université Des Frères Mentouri CONSTANTINE.

Statistique appliquée, Statistique des processus

Processus de Hawkes pour évaluer le risque de tavelure du pommier

Katarzyna Adamczyk-Chauvat [1] & Madalina Deaconu [2]

[1] *Université Paris-Saclay, INRAE, UR MaIAGE, France, Katarzyna.Adamczyk@inrae.fr*
[2] *Université de Lorraine, CNRS, Inria, IECL, France, Madalina.Deaconu@inria.fr*

Résumé. Le processus ponctuel de Hawkes modélise des phénomènes où l'occurrence d'un événement déclenche l'arrivée d'autres événements de même type. L'application présentée dans ce document concerne la tavelure, une maladie du pommier causée par un champignon pathogène *Venturia Inaequalis*. Le modèle de Hawkes proposé pour l'émission des spores du champignon tient compte de covariables météorologiques qui peuvent la déclencher. Les résultats montrent que l'estimateur de l'intensité conditionnelle du processus peut être considéré comme un indicateur pertinent du risque de contamination des pommiers.

Mots-clés. Processus de Hawkes, épidémiologie végétale, tavelure du pommier, covariables météorologiques.

Abstract. The Hawkes process models phenomena where the occurrence of an event triggers the arrival of other events of the same type. The application presented in this document concerns apple scab, a disease of apple trees caused by the pathogenic fungus *Venturia Inaequalis*. The Hawkes model proposed for the release of fungal spores takes into account meteorological covariates that may trigger it. The results show that the estimator of the conditional intensity of the process can be considered a relevant indicator of the risk of contamination of apple trees.

Keywords. Hawkes process, plant epidemiology, apple scab, meteorological covariates.

1 Introduction

Le processus ponctuel de Hawkes modélise des phénomènes où l'occurrence d'un événement déclenche l'arrivée d'autres événements de même type. Introduit par Alan G. Hawkes ([Hawkes, 1971], [Hawkes and Oakes, 1974]), le modèle a été appliqué dans diverses domaines : la modélisation des séismes, l'analyse financière ou encore l'étude de réseaux sociaux. Nous présentons un exemple d'application du processus de Hawkes en épidémiologie végétale, pour décrire la propagation de la tavelure, une maladie du pommier. La maladie est causée par un champignon pathogène, *Venturia Inaequalis*, qui se développe sur les feuilles infectées, tombées aux pieds des arbres en automne. Arrivant à maturité au début de printemps, le champignon libère des spores qui se dispersent et infectent de nouvelles feuilles et fruits. La lutte contre la maladie repose sur un traitement fongique qui doit être synchronisé avec l'émission des spores. Le modèle de Hawkes a pour objectif de prédire les instants d'émission des spores en fonction de conditions météorologiques.

2 Processus de Hawkes

On considère un processus ponctuel simple $(N(t))_{t\geq 0}$ sur une demi-droite. L'histoire du processus en temps t, notée $\mathcal{H}_t$, est définie comme l'ensemble d'évènements qui se sont produits avant l'instant

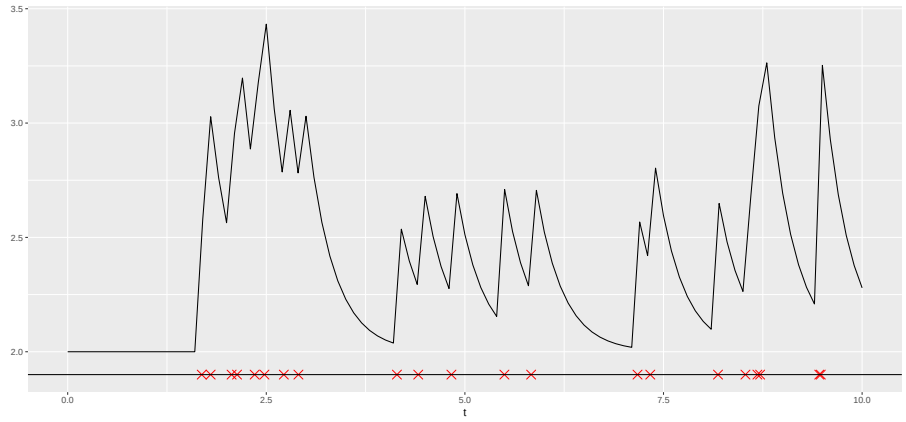


FIGURE 1 – Tracé de la fonction d'intensité conditionnelle $\lambda^\star(t)$ pour le processus de Hawkes avec $\lambda(t) \equiv 2$ et $\mu(t) = 0.6\exp(-3t)$ sur l'intervalle $]0, 10]$. Les croix représentent une trajectoire simulée du processus.

t. Le processus est caractérisé par la fonction d'intensité conditionnelle :

$$\lambda^\star(t) = \lim_{dt \to 0^+} \frac{\mathbb{E}\left(N(t+dt) - N(t) | \mathcal{H}_t\right)}{dt} \tag{1}$$

La valeur de $\lambda^\star(t)$ peut s'interpréter comme un risque instantané moyen de survenue d'un évènement à l'instant t, conditionnellement à l'histoire de processus avant t. On peut montrer que la fonction d'intensité conditionnelle (1) détermine la distribution du processus ponctuel de manière unique ([Daley and Vere-Jones, 2003], Proposition 7.2.IV).

Definition 2.1 (Processus de Hawkes)**.** Le processus de Hawkes est caractérisé par la fonction d'intensité conditionnelle :

$$\lambda^\star(t) = \lambda(t) + \sum_{t_i \in \mathcal{H}_t} \mu(t - t_i).$$

La fonction $\lambda(t) : \mathbb{R}^+ \to \mathbb{R}^+$ est appelé l'intensité de base et ne dépend pas de l'histoire du processus. La contribution des points du passé est donnée par la fonction d'excitation $\mu(t) : \mathbb{R}^+ \to \mathbb{R}^+$ telle que $\int_0^\infty \mu(s)ds < 1$.

La Figure 1 présente l'intensité conditionnelle de processus de Hawkes avec $\lambda(t) \equiv \lambda$ et le noyau exponentiel $\mu(t) = \alpha\exp(-\beta t)$, $0 < \alpha < \beta$. Dans cet exemple, les événements les plus récents ont une plus grande influence sur le processus. L'arrivée d'un nouveau évènement augmente l'intensité $\lambda^\star(t)$ de α et l'influence de cet évènement diminue avec le temps au taux β.

3 Inférence statistique

Diverses méthodes statistiques ont été développées pour inférer les paramètres de processus de Hawkes : ici nous nous focalisons sur l'estimation du maximum de vraisemblance. On considère une famille de processus ponctuels caractérisés par l'intensité $\lambda^\star_\theta(t)$ ou $\theta \in \mathbb{R}^d$ est un vecteur des paramètres. Soit $n(T)$ le nombre d'évènements observés dans $]0, T]$ et soit $0 < t_1, \ldots < t_{n(T)}$ une

trajectoire observée de processus. La fonction de log-vraisemblance du processus est donnée par :

$$\log L(\theta; \{t_i\}) = \sum_{i=1}^{n(T)} \log \lambda_\theta^\star(t_i) - \Lambda(T) \tag{2}$$

où $\Lambda(T) = \int_0^T \lambda_\theta^\star(t)dt$ est une fonction d'intensité cumulée. Ce résultat a été démontré dans [Daley and Vere-Jones, 2003]. Le comportement asymptotique de l'estimateur du maximum de vraisemblance $\hat{\theta}$ a été étudié dans [Ogata, 1978] pour les processus ponctuels stationnaires. La validation du modèle est basée sur le théorème (7.4.IV) de [Daley and Vere-Jones, 2003] :

Théorème 1 (Analyse résiduelle)**.** Soit $\{t_1, t_2, \ldots\}$ une trajectoire d'un processus ponctuel sur $(0, \infty)$ et soit $\Lambda(\cdot)$ la fonction d'intensité cumulée, supposée monotone, continue et telle que $\lim_{t\to\infty} \Lambda(t) = \infty$ p.s. La trajectoire transformée $\{t_1^\star, t_2^\star, \ldots\} = \{\Lambda(t_1), \Lambda(t_2), \ldots\}$ est une réalisation de processus de Poisson d'intensité 1 si et seulement si la trajectoire originale $\{t_1, t_2, \ldots\}$ est une réalisation de processus ponctuel défini par $\Lambda(\cdot)$.

Il s'ensuit que pour valider le modèle, il suffit de vérifier si la trajectoire transformée $\{t_1^\star, t_2^\star, \ldots\}$, restreinte à l'intervalle d'observations $]0, T]$, est une trajectoire de processus de Poisson unitaire. Pour ce faire, le test d'enveloppes globales basé sur des simulations de processus de Poisson unitaire peut être utilisé [Myllymäki et al., 2017].

4 Application aux données de la tavelure

Notre objectif est ici de prédire les instants d'émission de spores de *Venturia inaequalis*, le champignon à l'origine de la tavelure. L'émission des spores est déclenchée par les conditions météorologiques et continue pendant quelques heures qui suivent. L'expérience de monitoring des spores menée dans le verger expérimental de l'Institut National de Recherche en Horticulture à Skierniewice a fourni les données pour ajuster le modèle.

4.1 Capture des spores et données météorologiques

Un piège à spores a été placé sur le sol du verger expérimental recouvert de feuilles de la variété McIntosh fortement infectées par *V. Inaequalis* pendant la saison précédente [Jankowski and Masny, 2020]. L'appareil enregistrait l'information de présence-absence de spores à une fréquence horaire pendant deux mois suivant la date de détection des premières spores. En parallèle, la station météorologique placée sur le site prenait des mesures de variables météorologiques qui pouvaient déclencher l'émission des spores. La température (temp) et l'humidité de l'air (hAir), la pression atmosphérique (pres), la hauteur des précipitations (prec), l'humidité foliaire (hFol), la vitesse et la direction du vent (vitV, dirV), la luminosité (lumi) et l'intensité lumineuse (intL) font partie des variables collectées dans l'expérience. Ces variables, mesurées initialement à une échelle temporelle plus fine, ont été agrégées par heure pour correspondre aux données de spores.

4.2 Modèle d'émission des spores

Le processus de Hawkes proposé pour modéliser l'émission des spores est donné par l'intensité conditionnelle :

$$\begin{aligned} \lambda^\star(t) &= \lambda(t) + \alpha \sum_{t_i \in \mathcal{H}_t} \exp(-\beta(t - t_i)) \\ \log(\lambda(t)) &= \sum_{k=0}^{K} \gamma_k x_k(t) \end{aligned} \tag{3}$$

où $(x_1(t), \ldots, x_K(t))^T$ est un vecteur de K variables météorologiques mesurées au temps t et $x_0(t) \equiv 1$. L'intensité de base log-linéaire modélise l'impact des conditions climatiques sur le déclenchement de l'émission des spores et la partie auto-excitante tient compte de sa durée. Compte tenu des mesures expérimentales, l'intensité de base peut être approchée par la fonction constante par morceaux :

$$\log(\lambda(t)) = \sum_{k=0}^{K} \gamma_k \sum_{m=0}^{M-1} x_{k,m} \mathbb{1}_{(m,m+1]}(t)$$

où $]0, T] = \bigcup_{m=0}^{T-1}]m, m+1]$ et $x_{k,m}$ est la valeur de la k-ème variable sur l'intervalle de temps $]m, m+1]$. On assume par la suite que le processus a été observé jusqu'à la n-ème arrivée, on a alors $T = t_n$. La fonction de log-vraisemblance (2) pour le modèle (3) prend la forme :

$$\begin{aligned} \log L(\theta; \{t_i\}) &= \sum_{i=1}^{n} \log \left(\exp\left(\sum_{k=0}^{K} \gamma_k x_{k,m(i)} \right) + \alpha \sum_{j=1}^{i-1} \exp(-\beta(t_i - t_j)) \right) \\ &- \sum_{m=0}^{t_n - 1} \exp\left(\sum_{k=0}^{K} \gamma_k x_{k,m} \right) - \frac{\alpha}{\beta} \sum_{i=1}^{n-1} (1 - \exp(-\beta(t_n - t_i))) \end{aligned}$$

où $x_{k,m(i)}$ est la valeur de la k-ème variable observée au temps t_i, $i = 1, \ldots, n$, $k = 0, \ldots, K$. L'estimateur du maximum de vraisemblance de θ a été calculé à partir de données collectées en 2011. Cette année on a relevé 215 évènements d'émission des spores pendant 1416 heures de suivi. Les résultats d'estimation figurent dans la Table (1). Afin de comparer l'importance relative des covariables nous

cste	temp	hAir	pres	prec	hFol	vitV	dirV	lumi	intL		
$\hat{\gamma_0}$	$\hat{\gamma_1}$	$\hat{\gamma_2}$	$\hat{\gamma_3}$	$\hat{\gamma_4}$	$\hat{\gamma_5}$	$\hat{\gamma_6}$	$\hat{\gamma_7}$	$\hat{\gamma_8}$	$\hat{\gamma_9}$	$\hat{\alpha}$	$\hat{\beta}$
-4.26	0.40	1.06	-0.49	0.03	0.91	-0.11	0.06	0.59	-0.08	0.21	0.36

TABLE 1 – Estimation du maximum de vraisemblance pour le modèle 3. Les paramètres γ_i correspondent aux variables météorologiques, les paramètres α et β apparaissent dans la fonction d'excitation du modèle.

avons utilisé l'information sur l'ordre de leur introduction dans le modèle. Nous avons appliqué la stratégie ascendante, en ajoutant à chaque étape la variable qui améliore le plus ou détériore le moins l'AIC (Figure 2). La variable qui se place en tête du classement c'est l'humidité foliaire. Elle est suivie par la luminosité, l'humidité de l'air, la pression atmosphérique et la température : ces variables font cependant décroître l'AIC moins rapidement. L'ajout d'autres variables météorologiques fait augmenter le critère.

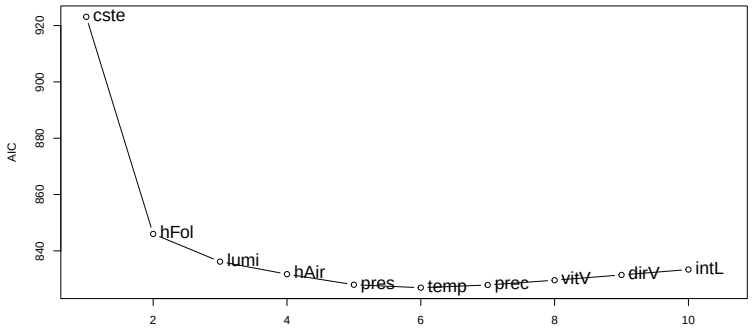


FIGURE 2 – L'évolution de l'AIC lors de l'ajout des variables météorologiques au modèle. A chaque étape on choisit la variable qui améliore le plus ou détériore le moins l'AIC, en commençant par le modèle avec l'intensité de base constante.

Le test d'enveloppes globales basé sur 500 simulations de processus de Poisson d'intensité 1 ne rejette pas l'hypothèse nulle de compatibilité entre modèle et données avec la p-valeur de 0.63. La Figure 3 présente l'interprétation graphique du résultat de test : le comptage associé à la trajectoire transformée $\{t_1^\star, t_2^\star, \ldots, t_n^\star\}$ reste à l'intérieur de la région délimitée par la procédure de test. Par conséquent la trajectoire originale $t_1, \ldots, t_n$ peut être modélisée comme une réalisation du modèle 3.

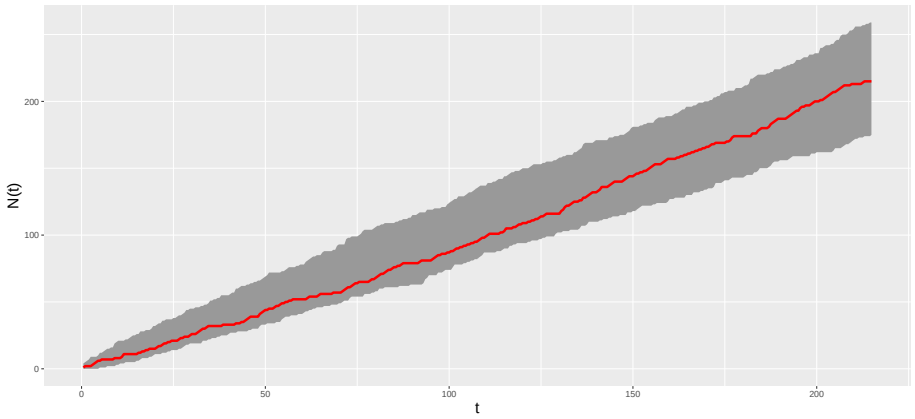


FIGURE 3 – Test d'enveloppes globales pour l'analyse résiduelle du modèle de Hawkes 3. La courbe rouge représente le comptage associé à la trajectoire de processus transformé, $\{\Lambda_{\hat{\theta}}(t_1), \Lambda_{\hat{\theta}}(t_2), \ldots, \Lambda_{\hat{\theta}}(t_n)\}$. La courbe est située à l'intérieur de la région délimitée par le test, marquée en gris.

.

5 Conclusions

Ce document présente l'application de processus de Hawkes à la modélisation de l'émission des spores de *V. Inaequalis*. L'intensité conditionnelle du modèle proposé a été construite en utilisant la fonction d'excitation à décroissance exponentielle et l'intensité de base dépendant de covariables météorologiques. La méthode du maximum de vraisemblance a été utilisée pour estimer les paramètres du modèle. La comparaison de variables météorologiques a mis en évidence le rôle de variables liées à l'humidité dans le déclenchement de l'émission des spores. Le test d'enveloppes globales utilisant l'analyse résiduelle a été appliqué pour vérifier la qualité d'ajustement du modèle. Les résultats confirment que l'estimateur de l'intensité conditionnelle du modèle peut être utilisé comme un indicateur de risque de contamination par le champignon.

Références

[Daley and Vere-Jones, 2003] Daley, D. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes, Second Edition*. Springer, New York.

[Hawkes, 1971] Hawkes, A. (1971). Spectra of some self-exciting and mutually-exciting point processes. *Biometrika*, 58 :83–90.

[Hawkes and Oakes, 1974] Hawkes, A. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11 :493–503.

[Jankowski and Masny, 2020] Jankowski, P. and Masny, S. (2020). Influence of moisture on maturation rate of the *Venturia inaequalis* (cooke) wint. ascospores in central poland. *Journal of Plant Diseases and Protection*, 127 :155–163.

[Myllymäki et al., 2017] Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., and Hahn, U. (2017). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 79 :381–404.

[Ogata, 1978] Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Ann. Inst. Statist. Math.*, 30 :243–261.

[Pierrat, 2022] Pierrat, P. (2022). Processus de Hawkes : application en épidémiologie végétale. Technical report, INRAE - Département Mathématiques et Numérique, Inria - Équipe PASTA.

Statistique Bayésienne (Session du groupe)

Learnings from a large simulation study to optimize Bayesian trials in small populations

Billy Amzal*[1]

[1]Quinten Health – Quinten Health – France

Résumé

Extrapolating treatment effects from related studies is a promising strategy for designing and analyzing clinical trials in situations where achieving an adequate sample size is challenging. Bayesian methods are well-suited for this purpose, as they enable the synthesis of prior information through the use of prior distributions. While the operating characteristics of Bayesian approaches for borrowing data from control arms have been extensively studied (Viele et al. 2014), methods that borrow treatment effects-quantities derived from the comparison between two arms-remain less well understood. In this paper, we present how an extensive and a computation-intensive simulation study can be designed to address this gap. We evaluate the frequentist operating characteristics of these methods, including the probability of success, mean squared error, bias, precision, and credible interval coverage. Our results provide insights into the strengths and limitations of existing methods in the context of confirmatory trials. In particular, we show that the Conditional Power Prior and the Robust Mixture Prior perform better overall, while the test-then-pool variants and the p-value-based power prior display suboptimal performance.

Mots-Clés: Pediatric extrapolation, Bayesian trial designs, simulation study

*Intervenant

Estimating ratios of normalizing constants using stochastic approximation: the SARIS algorithm

Tom Guédon [1] & Charlotte Baey [2] & Estelle Kuhn [1]

[1] *Université Paris-Saclay, INRAE, MaIAGE, Domaine de Vilvert, Jouy-en-Josas, 78350, France*

[2] *Laboratoire Paul Painlevé, Université de Lille, Cité scientifique, Villeneuve-d'Ascq, 59655, France*

Résumé. Le calcul des rapports des constantes de normalisation joue un rôle important dans la modélisation statistique. Deux exemples importants sont les tests d'hypothèses dans les modèles à variables latentes et la comparaison de modèles en statistique bayésienne. Dans les deux examples, elles s'écrivent comme un rapport de constantes de normalisation : pour les modèles à variables latentes, celles des distributions conditionnelles aux variables observées, et en statistique bayésienne, celles des distributions a posteriori. Nous proposons dans cet article une nouvelle méthodologie qui estime ce rapport en utilisant le principe d'approximation stochastique. Notre estimateur est consistant et asymptotiquement gaussien. Sa variance asymptotique est plus petite que celle de l'estimateur optimal par bridge sampling. De plus, il est plus robuste à un faible chevauchement des supports des deux distributions non normalisées considérées. Grâce à sa définition itérative, notre procédure peut être intégrée dans un processus d'estimation dans les modèles à variables latentes, réduisant ainsi l'effort de calcul. Les performances de l'estimateur sont illustrées par une étude de simulation et comparées à deux autres estimateurs : le ratio importance sampling et le bridge sampling.

Mots-clés. ratio de constantes de normalisation, méthodes Monte Carlo, Approximation stochastique, estimation de vraisemblance marginale

Abstract. Computing ratios of normalizing constants plays an important role in statistical modeling. Two important examples are hypothesis testing in latent variables models, and model comparison in Bayesian statistics. In both cases, they can be written as a ratio of normalization constants: for latent variable models, those of the conditional distributions given the observed data, and in Bayesian statistics, those of the posterior distributions. We propose in this article a new methodology that estimates this ratio using stochastic approximation principle. Our estimator is consistent and asymptotically Gaussian. Its asymptotic variance is smaller than the one of the popular optimal bridge sampling estimator. Furthermore, it is much more robust to little overlap between the two unnormalized distributions considered. Thanks to its online definition, our procedure can be integrated in an estimation process in latent variables model, reducing the computational effort. The performances of the estimator are illustrated through a simulation study and compared to two other estimators : the ratio importance sampling and the optimal bridge sampling estimators.

Keywords. ratio of normalizing constants, Monte Carlo methods, Stochastic approximation, marginal likelihood estimation

This talk is based on the paper [5]

1 Statistical setting and objective

Let d be a positive integer and μ a σ-finite positive Borel measure on a subspace $\mathcal{Z}$ of $\mathbb{R}^d$. Assume that f_0 and f_1 are two positive integrable Borel functions on $\mathcal{Z}$ such that $c_0 = \int_{\mathcal{Z}} f_0(z)\mu(dz) > 0$ and $c_1 = \int_{\mathcal{Z}} f_1(z)\mu(dz) > 0$. We assume that these normalizing constants c_0 and c_1 are unknown and introduce the two probability densities with respect to μ denoted by p_0 and p_1 defined for $i \in \{0, 1\}$ and for all z in $\mathcal{Z}$ by

$$p_i(z) = \frac{f_i(z)}{c_i}.$$

The objective is to estimate the ratio r^* of normalizing constants defined as:

$$r^* = \frac{c_0}{c_1} = \frac{\int_{\mathcal{Z}} f_0(z)\mu(dz)}{\int_{\mathcal{Z}} f_1(z)\mu(dz)}. \tag{1}$$

We first motivate our contribution by two practical examples which require the computation of such ratios.

Example 1 (**computation of likelihood ratio test statistic in latent variables model**). *Let us consider a general latent variables model where the observed variable is given by the random variable Y, taking values in $\mathcal{Y}$ and the latent variable by Z, taking values in $\mathcal{Z}$. We denote by y the observation of Y. We assume that the random vector (Y, Z) follows a parametric distribution with density f_θ parameterized by $\theta \in \Theta$. The objective is to test the hypotheses:*

$$H_0 : \quad \theta \in \Theta_0 \quad \textit{against} \quad H_1 : \quad \theta \in \Theta_1,$$

where $\Theta_0 \subset \Theta_1 \subset \Theta$. A natural popular test is the likelihood ratio test [9] which statistic is defined by:

$$LR = -2 \log \left(\frac{L(\hat{\theta}_0; y)}{L(\hat{\theta}_1; y)} \right),$$

where the marginal likelihood $L(\theta; y)$ and the maximum likelihood estimates $\hat{\theta}_i$, for $i \in \{0, 1\}$, are defined respectively by

$$L(\theta; y) = \int_{\mathcal{Z}} f_\theta(y; z)\mu(dz)$$

and

$$\hat{\theta}_i = \arg\max_{\theta \in \Theta_i} \quad L(\theta; y).$$

Accurately estimating LR is crucial as its value determines whether H_0 is rejected or not.

Example 2 (**computation of Bayes factor for Bayesian model choice**). *In Bayesian statistics, the parameter $\theta \in \Theta$ is considered as a random variable with a known prior distribution $p(\theta)$. The posterior distribution of θ given a dataset D is defined as the product of*

the prior density and the likelihood of the model up to a normalizing constant. To compare two models M_1 and M_2 and choose the one that better fits the data, the Bayes factor B_{12} [4] is a powerful tool. It is defined as:

$$B_{12} = \frac{p(D \mid M_1)}{p(D \mid M_2)}$$

where $p(D \mid M_i) = \int_\Theta p(D \mid \theta, M_1)p(\theta)d\theta$ is the marginal likelihood of model M_i, for $i \in \{1, 2\}$.

2 Description of the SARIS algorithm

Let π be a positive density function on $\mathcal{Z}$. Starting from equation (1), rewritten as:

$$\int_{\mathcal{Z}} (f_0(z) - r^* f_1(z))\, \mu(dz) = 0, \tag{2}$$

we can write:

$$\mathrm{E}_\pi \left[\frac{f_0(Z) - r^* f_1(Z)}{\pi(Z)} \right] = 0, \tag{3}$$

where E_π denotes expectation with respect to the density π.

This formulation corresponds to a root-finding problem involving an expectation, and thus naturally fits into the Robbins-Monro stochastic approximation framework [7]. Assuming that $r_0 \in \mathbb{R}$ is given and that one can sample independent draws from π, we consider the sequence $(r_k)_{k\geq 0}$ of estimators of r^* defined by the following recursion:

$$r_{k+1} = r_k + \gamma_{k+1} \frac{f_0(Z_{k+1}) - r_k f_1(Z_{k+1})}{\pi(Z_{k+1})}, \quad \text{with} \quad Z_{k+1} \sim \pi, \tag{4}$$

where $(\gamma_k)_{k\geq 0}$ is a sequence of positive decreasing step sizes.

The main task in the construction of the sequence is the choice of the proposal distribution π. Most relevant choices for π may depend on the true ratio r^*. For example, the optimal choice for Bridge sampling involves the quantity $p_0 + p_1$, which itself depends on r^* and is therefore not directly accessible. This dependency is also a main drawback of optimal ratio importance sampling.

Thanks to its iterative nature, our methodology allows for proposal distributions that depend on r^*. More precisely, let us consider a family of positive densities π_r on $\mathcal{Z}$ indexed by r. Equation (3) can then be reformulated as:

$$\mathrm{E}_{\pi_{r^*}} \left[\frac{f_0(Z) - r^* f_1(Z)}{\pi_{r^*}(Z)} \right] = 0. \tag{5}$$

This can again be tackled using the Robbins-Monro algorithm, now in an adaptive form, through the stochastic recursion:

$$r_{k+1} = r_k + \gamma_{k+1} \frac{f_0(Z_{k+1}) - r_k f_1(Z_{k+1})}{\pi_{r_k}(Z_{k+1})}, \quad \text{with} \quad Z_{k+1} \sim \pi_{r_k}. \tag{6}$$

The general algorithm is called SARIS (Stochastic Approximation Ratio Importance Sampling) and is summarized in Algorithm 1.

Algorithm 1 SARIS algorithm

Input: $(\gamma_k)_{k\geq 0}$, r_0, stopping criterion
Until stopping criterion:
  Draw Z_{k+1} from π_{r_k}
  Update $r_{k+1} = r_k + \gamma_k \frac{f_0(Z_{k+1}) - r_k f_1(Z_{k+1})}{\pi_{r_k}(Z_{k+1})}$
  $k = k + 1$
Return r_k

Remark 1. *In most real-life applications, it is difficult to independently and exactly simulate from complex distributions. Therefore, the simulation step in* (6) *might be intractable. It is however possible to use the transition kernel of an ergodic Markov Chain having* π_{r_k} *as invariant distribution. One common practical choice for such Markov Chain Monte Carlo (MCMC) sampling scheme is the Metropolis-Hastings or the Metropolis-within-Gibbs algorithm [8].*

Remark 2. *We emphasize that the SARIS algorithm can also be used to compute a single marginal likelihood. If we know a density* p *only up to a normalizing constant* c, $p = f/c$, *then by introducing a known normalized density* g, *the SARIS algorithm can be used to compute the ratio* $r^* = c$.

Remark 3. *Finally, the proposed procedure enables to directly estimate any strictly monotonous and invertible transformation* g *of the ratio. Suppose that the objective is to compute* $g(r^*)$ *(for example* $g = -2\log$ *to obtain a likelihood ratio statistic). The recursive scheme* (6) *can be easily modified to estimate* $g(r^*)$ *with the sequence* $(g_k)_{k\geq 0}$ *defined as follows:*

$$g_{k+1} = g_k + \gamma_{k+1} \frac{f_0(Z_{k+1}) - g^{-1}(g_k) f_1(Z_{k+1})}{\pi_{g^{-1}(g_k)}(Z_{k+1})} \qquad Z_{k+1} \sim \pi_{g^{-1}(g_k)} \tag{7}$$

Algorithm 1 can be easily adapted using this recursion, only changing the updating rule.

It is important to notice that the algorithm presented in this section is not applicable when the analytical expression of the proposal density π_r is unknown, and therefore the update rule of the sequence r_k defined in Algorithm 1 is not computable. The next section solves this issue.

2.1 A practical extension of the SARIS algorithm

Suppose that for all z, $\pi_r(z) = \tilde{\pi}_r(z)/c(r)$ where the analytical expression of $\tilde{\pi}_r(z)$ is known, then equation (3) is equivalent to:

$$\mathrm{E}_{\pi_{r^*}}\left[\frac{f_0(Z) - r^* f_1(Z)}{\tilde{\pi}_{r^*}(Z)}\right] = 0. \tag{8}$$

Based on this identity, one can consider the following extension of Algorithm 1 called SARIS-EXT, that uses the known unnormalised density $\tilde{\pi}_r$:

Algorithm 2 SARIS-EXT algorithm

Input: $(\gamma_k)_{k\geq 0}$, r_0, stopping criterion
Until stopping criterion:
Draw Z_{k+1} from $\pi_{r_k} \propto \tilde{\pi}_{r_k}$
Update $r_{k+1} = r_k + \gamma_k \frac{f_0(Z_{k+1}) - r_k f_1(Z_{k+1})}{\tilde{\pi}_{r_k}(Z_{k+1})}$
$k = k + 1$
Return r_k

This algorithm presents the same asymptotic performances as the one of algorithm 1. For theoretical properties of the procedures, see [5].

3 Theoretical property of the SARIS algorithm

In this section we study the theoretical convergence property of the sequence $(r_k)_{k\geq 0}$ generated by the SARIS procedure described in Algorithm 1.

We first state some regularity assumptions on the functions f_0, f_1 and on the densities $\{\pi_r, r \in \mathbb{R}\}$:

Assumption 1. *The functions f_0 and f_1 are positive integrable and for every $z \in \mathcal{Z}$, $r \mapsto \pi_r(z)$ is continuous.*

Assumption 2.

$$\mathrm{E}_0\left(\sup_{r\in\mathbb{R}}\left|\frac{f_0(Z) - r f_1(Z)}{\pi_r(Z)}\right|\right) + \mathrm{E}_1\left(\sup_{r\in\mathbb{R}}\left|\frac{f_0(Z) - r f_1(Z)}{\pi_r(Z)}\right|\right) < +\infty.$$

This assumption ensures the integrability of the main quantities involved in the algorithm. We also state a common assumption on the sequence of step sizes $(\gamma_k)_{k\geq 0}$.

Assumption 3. *The sequence $(\gamma_k)_{k\geq 0}$ is positive, decreasing and verifies $\sum_{k=0}^{+\infty} \gamma_k = +\infty$ and $\sum_{k=0}^{+\infty} \gamma_k^2 < +\infty$.*

We can now state the almost sure (a.s.) convergence of the sequence (r_k).

Proposition 1. *Considering the sequence $(r_k)_{k\geq 0}$ generated by Algorithm 1, under Assumptions 1, 2 and 3, we get:*

$$\lim_{k\to+\infty} \quad r_k = r^* \quad a.s.$$

See [5] for more details. We now require a stronger regularity assumption on the functions f_0, f_1, π_r to derive the asymptotic distribution of the sequence $(r_k)_{k\geq 0}$.

Assumption 4. *The functions f_0 and f_1 are not proportional to each other.*

Assumption 5. *There exists $\delta > 0$ such that*

$$\sup_{k\geq 0} \quad \mathrm{E}_0\left(\left|\frac{f_0(Z) - r_k f_1(Z)}{\pi_{r_k}(Z)}\right|^{1+\delta}\right) + \sup_{k\geq 0} \quad \mathrm{E}_1\left(\left|\frac{f_0(Z) - r_k f_1(Z)}{\pi_{r_k}(Z)}\right|^{1+\delta}\right) < +\infty.$$

Assumption 6. *There exists $\frac{1}{2} < \epsilon < 1$, $a > 0$, $b > 0$ such that the sequence of step sizes (γ_k) is of the form $\gamma_k = \frac{a}{b+k^\epsilon}$.*

These integrability and step sizes assumptions are classical ones to obtain asymptotic normality results for martingales. The next result states the asymptotic normality of the averaged sequence defined as $(r_k^{AV})_{k\geq 0} = \left(\frac{1}{k}\sum_{j=0}^{k} r_j\right)_{k\geq 0}$.

Proposition 2. *Considering the sequence $(r_k)_{k\geq 0}$ generated by Algorithm 1 and its averaged version $(r_k^{AV})_{k\geq 0}$, under Assumptions 1, 4, 5 and 6, we get:*

$$\sqrt{k}\left(r_k^{AV} - r^*\right) \xrightarrow[k\to+\infty]{d} \mathcal{N}\left(0, V_{saris}(\pi_{r^*})\right)$$

with

$$V_{saris}(\pi_{r^*}) = \frac{1}{{c_1}^2}\mathrm{E}_{r^*}\left[\left(\frac{f_0(Z) - r^* f_1(Z)}{\pi_{r^*}(Z)}\right)^2\right]$$

where the expectation E_{r^*} is taken with respect to the density π_{r^*}. Furthermore, the optimal proposal π_{saris}^{opt} defined as the one which minimizes the asymptotic variance is given as

$$\pi_{r^*}^{opt}(z) \propto |p_1(z) - p_0(z)|,$$

corresponding to the optimal variance:

$$V_{saris}^{opt} = r^{*2}\left(\int_{\mathcal{Z}} |p_1(z) - p_0(z)|\mu(dz)\right)^2.$$

Remark 4. *If sampling is done through the use of a Markov transition kernel, similar theoretical results can be obtained assuming additional regularity conditions on the Markov kernel. For further details we refer to [1] and [3].*

In [5] it is shown that the same asymptotic results are achieved for the SARIS-EXT algorithm.

4 Synthesis of the contribution

We proposed a new methodology to compute ratios of normalizing constants that relies on the stochastic approximation principle. In [5], this procedure is compared to the popular bridge sampling (BS) of [6] and the ratio importance sampling (RIS) of [2]. Theoretically

SARIS and RIS have the same asymptotic optimal variance, bounded, that is smaller than the one of the optimal BS (unbounded). However the optimal scheme of SARIS is computable contrary to the one of RIS. Figure(4) compares the optimal BS estimator and the optimal SARIS estimator, in a Gaussian example with p_0 being the density of $\mathcal{N}(0,1)$, and p_1 the one of a $\mathcal{N}(\mu,1)$. This simple example illustrates the fact that even when the overlap between p_0 and p_1 decreases, SARIS still manage to compute r^*, contrary to the optimal BS estimator.

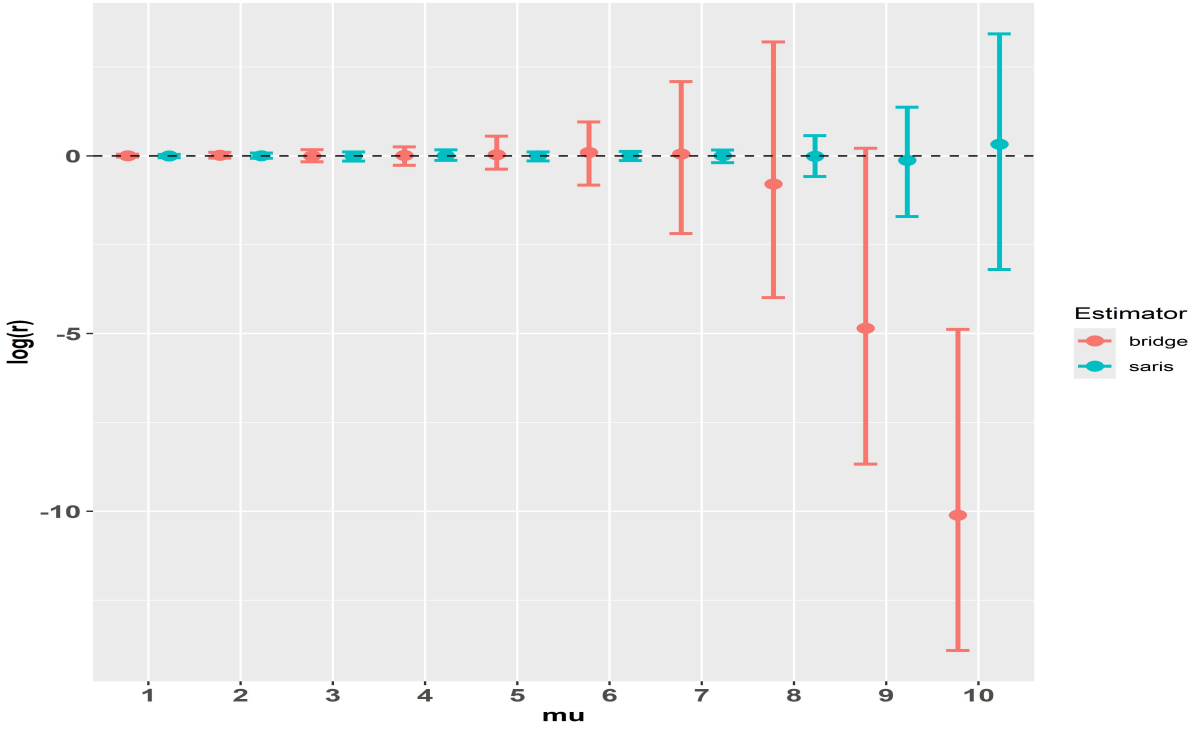


Figure 1: Estimation of the log ratio of normalizing constants of the densities of a $\mathcal{N}(0,1)$ and the one of a $\mathcal{N}(\mu,1)$ for varying values of μ, using the optimal Bridge sampling estimator (red) and the optimal SARIS estimator (blue) to compute the ratio. The dots represent the empirical means and the error bars the 5% and 95% empirical quantiles computed over 50 repetitions. For each experiment, the same number of evaluations of the densities were used. Every simulations was performed using an adaptive Metropolis-Hasting algorithm

References

[1] Stéphanie Allassonniere and Estelle Kuhn. Convergent stochastic expectation maximization algorithm with efficient sampling in high dimension. application to deformable template model estimation. *Computational Statistics & Data Analysis*, 91:4–19, 2015.

[2] Ming-Hui Chen and Qi-Man Shao. On monte carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25(4):1563–1594, 1997.

[3] Gersende Fort. Central limit theorems for stochastic approximation with controlled markov chain dynamics. *ESAIM: Probability and Statistics*, 19:60–80, 2015.

[4] Alan E Gelfand and Dipak K Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, 1994.

[5] Tom Guédon, Charlotte Baey, and Estelle Kuhn. Estimation of ratios of normalizing constants using stochastic approximation: the saris algorithm. *arXiv preprint arXiv:2408.13022*, 2024.

[6] Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

[7] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[8] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, New York, NY, 1999.

[9] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, Cambridge, 2000.

Calage bayésien avec entrées incertaines, application à un modèle de sas de dépressurisation

Edimah Synesius Songo[1] & Joseph Muré[1] & Merlin Keller[1]

[1] *EDF, Recherche & Développement*
Dpt. Performance, Risque Industriel, Surveillance
pour la Maintenance et l'Exploitation
6 quai Watier, 78400 Chatou, France

Résumé Les simulations numériques sont couramment utilisées dans les études industrielles, en complément ou en remplacement d'expériences physiques coûteuses. Elles sont basées sur des modèles mathématiques dont les paramètres incertains doivent être estimés sur la base des données de terrain disponibles. Cette tâche, connue sous le nom de calage, peut être considérée comme un cas particulier de régression non linéaire.

Les méthodes de calage les plus répandues, telles que la méthode des moindres carrés ordinaires, ignorent généralement l'incertitude sur les variables en entrée du modèle, bien qu'elles soient entâchées d'erreurs de mesure de la même manière que les sorties.

Dans ce travail, nous comparons plusieurs méthodes de calage issues de la littérature, qui traitent des entrées incertaines d'un point de vue bayésien, et nous les appliquons à un modèle de débit d'air de sas dépressurisés présents dans les centrales nucléaires. En particulier, nous introduisons une nouvelle stratégie, basée sur la quantisation de la distribution *a priori* des données d'entrée, qui permet de traiter des données en nombre important.

Mots-clefs Calage bayésien, Régression avec erreurs sur les variables, quantisation, MCMC

Abstract Numerical simulations are routinely used in industrial studies, to complement or replace costly physical experiments. Such simulations are based on mathematical models whose uncertain parameters need to be estimated based on available field data. This task, known as calibration, can be seen as a special case of nonlinear regression.

The most widespread calibration methods, such as ordinary least-squares, typically ignore the uncertainty tainting the observed covariates, even though they are plagued with measurement errors in the same manner as the observed outcome.

In this work, we compare several calibration methods from the literature that deal with uncertain inputs from a Bayesian perspective, and apply them to an airflow rate model of depressurized rooms in nuclear power plants. In particular, we introduce a novel strategy, based on quantization of the inputs joint prior distribution, allowing to deal with large datasets.

Keywords Bayesian calibration, regression with error in variables, quantisation, MCMC

1 Introduction

Numerical simulations are routinely used in industrial studies to complement or replace costly physical experiments. Such simulations are based on mathematical models which predict a certain outcome variable y as a function g of controled covariates x and uncertain parameters θ, which are usually interpreted as more or less well-known physical constants. These must be estimated from available field data in the form of observed pairs $(x_i, y_i)_{1\leq i\leq n}$. This task, known as calibration, can be seen as a special case of nonlinear regression. In the industrial practice, Bayesian approaches are popular since they allow to deal with poorly informative data, and incorporate expert information under the form of prior laws [1].

However, calibration methods typically ignore the uncertainty tainting the observed covariates x_i, even though they are plagued by measurement errors in the same way as the observed outcome y_i. [2] deals with this issue by considering each uncertain covariate x_i as an additional calibration parameter. From a Bayesian perspective, this leads to a potentially high-dimensional posterior which can be hard to sample or otherwise determine. Still in [2], a preliminary Monte Carlo integration over the joint prior of the outcomes x_i is used to simplify the task.

In this paper, we compare the above strategies by applying them to an industrial case study involving a very simple numerical model, using a rather large dataset of several thousands of observations.

This requires implementing specific Monte Carlo Markov Chain (MCMC) algorithms [3] to sample from the posterior law of the calibration parameters. We use to this end the OpenTURNS software [4], developed by a consortium involving EDF among ther industrial partners, which includes classes dedicated to MCMC.

The following of the paper is organized as follows. In Section 2, we start by presenting the case-study. In Section 3 we apply a basic Bayesian calibration method (ignoring input uncertainty). In Section 4, we use joint Bayesian calibration extended to the uncertain inputs, while in Section 5 a novel approach is introduced, based on a preliminary numerical integration of the uncertain inputs after *quantization* (approximation by a discrete law) of their joint distribution. After comparing results from these different strategies, we end in Section 6 by a discussion on the perspectives opened by this work.

2 Industrial case study

In nuclear safety, depressurised rooms (or vacuum locks) are crucial to isolate radioactive particles, and very straightforward: a tube is used to vacuum the air out of the room. The airflow rate generated in the process can be controlled to guarantee the creation of the depression.

Said rate y can be modeled by a function of the pressure difference, noted x in the following, and parameterised by $\theta = (\theta_0, \theta_1) \in \mathbb{R}_+^2$, with θ_0 the total leakage area, and θ_1 an exponent parameter representing whether the leakage is quadratic:

$$g(x, \theta) = 0.6 \times 3600 \times \theta_0 \left(\frac{2}{1.8} x \right)^{\theta_1}$$

The constant coefficients of this code are conversion units and constants obtained with the application of a fluid mechanic principle called the Darcy-Weisbach equation.

We received a set of $n = 2\,336$ measurements of pressure differences in bar, $\mathbf{x} := (x_i)_{i \in \{1,\dots,n\}}$, and airflow rates in m^3/h, $\mathbf{y} := (y_i)_{i \in \{1,\dots,n\}}$. The associated dataset $\mathcal{D} = [\mathbf{x}, \mathbf{y}]$ is a non-publicly available CSV file based on EDF's own internal measurements. In our specific dataset, pressure measurements are rounded to a tenth, which is a confirmed source of input uncertainty.

3 Bayesian calibration with no input uncertainty

The most popular approach to code calibration considers only output uncertainty, following:

$$y_i \;=\; g(x_i, \theta) + E_i \quad \text{with } E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_E^2),$$

where E_i models flow rate measurement errors, while the pressure difference measures x_i are considered exact.

3.1 Prior density

We fix the following 2-dimensional prior for θ :

$$[\theta] \sim \mathcal{U}([0,2]) \otimes \mathcal{U}([0,2])$$

The bounds were provided by experts of air flow models, as containing all physically plausible values for the model parameters.

Meanwhile, σ_E is endowed with the minimally informative improper Jeffreys prior:

$$[\sigma_E] \propto \sigma_E^{-1},$$

as is usual in the absence of prior information.

3.2 Posterior density

Following [5], we choose to integrate out the output standard error σ_E^2 analytically from the joint posterior density expression. The resulting posterior is then given by :

$$\begin{aligned} [\theta|\mathbf{x}, \mathbf{y}] &\propto [\mathbf{y}|\theta, \mathbf{x}][\theta] \\ &\propto ||\mathbf{y} - \mathbf{g}(\mathbf{x}, \theta)||^{-\frac{n}{2}} \mathbf{1}_{[0,2]\times[0,2]}(\theta) \end{aligned}$$

We then use the above distribution, known up to a multiplicative constant, as the target of a random walk Metropolis-within-Gibbs algorithm using Gaussian proposal kernels.

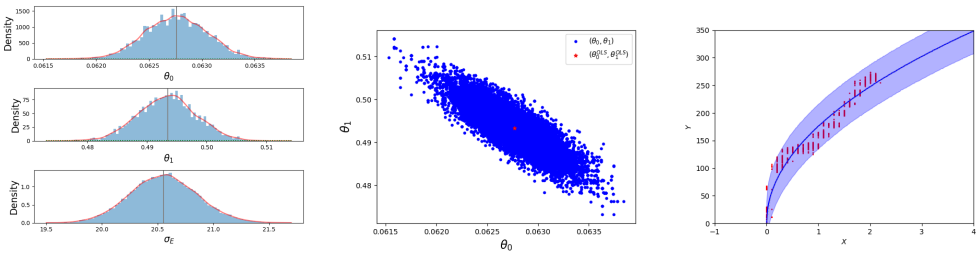


Figure 1: Result of Basic Bayesian Calibration. Left: Marginal posterior distributions of θ_0 (top), θ_1 (middle) and σ_E (bottom); the vertical line corresponds to the posterior median. Center: Scatter plot of the sampled θ values (blue dots), alongside the OLS estimator (red star). Right: Posterior 95% prediction intervals (blue) against obervations (red points).

3.3 Results

we ran 3 corresponding parallel chains of size 10 000. The random-walk Metropolis-Hastings algorithm in OpenTURNS includes a burn-in period, whose default value is set to 1 000, during which the standard deviations of said kernels are automatically tuned to attain acceptance rates between 20% and 80%, promoting the mixing and convergence of the chains to their equilibrium distribution. The Gelman-Rubin [6] convergence test results were approximately 1 for each component, and the Auto-Correlation Function (ACF) showed adequate mixing.

Having removed the first 1 000 iterations from the three Markov chains, we pooled the remaining 27 000 iterations. The resulting joint posterior and predictive distributions are displayed in Figure 1. Although convergence occurs, the ensuing 95% credibility region does not account for input uncertainty, and cannot be guaranteed to provide reliable predictions. The overall run time of this first analysis took 45 seconds (see Table 1).

4 Bayesian calibration with input uncertainty

Following [2], we now account for input uncertainty, by introducing the additional nuisance parameters ξ_i, corresponding to the inputs actual value, of which only a noisy measure x_i is available. In our case, this leads to the

following hierarchical model:

$$\begin{aligned} y_i &= g(\xi_i, \theta) + E_i \quad E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_E^2) \\ \xi_i &= x_i + Z_i \qquad Z_i \overset{ind}{\sim} \mathcal{U}([\max(-\epsilon, -x_i), \epsilon]). \end{aligned}$$

where $\epsilon = 5.10^{-2}$ represents the uniform error due to round-off, interpreted here as a prior distribution. Note that Z_i is additionally constrained not to be lower than $-x_i$, to prevent $\xi_i < 0$.

4.1 Posterior density

Our new target distribution is $[\theta, \xi | \mathbf{x}, \mathbf{y}]$. We integrate as before σ_E^2 from the likelihood, leading to the following posterior distribution:

$$[\theta, \boldsymbol{\xi} | \mathbf{x}, \mathbf{y}] \propto ||\mathbf{y} - \mathbf{g}(\boldsymbol{\xi}, \theta)||^{-\frac{n}{2}} \times \mathbf{1}_{[0,2]\times[0,2]}(\theta) \times \prod_{i=1}^{n} \mathbf{1}_{[\max(0, x_i - \epsilon), x_i + \epsilon]}(\xi_i).$$

As previously, a component-wise Metropolis-within-Gibbs implementation is a natural choice to sample θ. For the inputs, we used the marginals of the uniform prior distribution on $\boldsymbol{\xi}$ as proposition kernels to sample each individual ξ in an Independent Metropolis-Hastings [3] approach.

4.2 Results

We ran three parallel chains with random prior initialisations. This took about 5 hours and 20 minutes (see Table 1). The acceptance rates for θ were in the acceptable 20-80% range. On the other hand, acceptance rates for $\boldsymbol{\xi}$ ranged from 2% to 27%, which is considered small for Independent Metropolis-Hastings sampling, where we try to attain rates close to 100%. However, convergence and adequate mixing of the chains were confirmed by both visual tests and the Gelman Rubin criterion.

The joint posterior densities for θ are shown in Figure 2. With respect to the Basic Bayesian Calibration approach, the main difference lies in the estimation of the measurement error standard deviation σ_E, which is reduced by approximately 15%. This translates in tighter prediction bounds, while the estimated calibration parameters remain largely unchanged.

This confirms the advantage of taking into account input uncertainty in our case, however this comes at a significant increase in the computational cost, in our case from less than a minute to several hours.

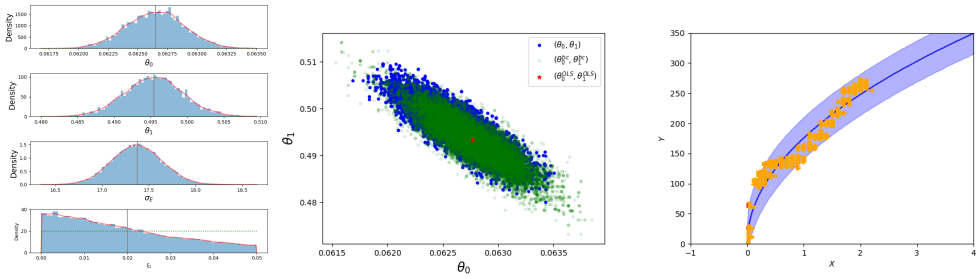


Figure 2: of joint calibration. Left: Marginal posterior distributions of θ_0, θ_1, σ_E and ξ_1 ; the vertical line corresponds to the posterior median. Center: Scatter plot of the sampled θ values (blue dots), alongside the OLS estimator (red star) and the posterior density without input uncertainty (Basic Bayesian calibration, green dots). Right: Posterior 95% prediction intervals (blue) against obervations (orange error bars showing the posterior uncertainty on the inputs).

5 Numerical integration over the input space

As shown in [2], an alternative to direct joint sampling of θ and $\boldsymbol{\xi}$ consists in reducing the dimension of the Markov chain by pre-integrating over the latent variables $\boldsymbol{\xi}$.

5.1 Quantized input prior

To write the corresponding target, we start with the expression of the joint posterior for (θ, σ_E, ξ), before integrating σ_E out, yielding the following factorization:

$$[\theta, \sigma_E, \boldsymbol{\xi}|\mathbf{x}, \mathbf{y}] \propto \prod_{i=1}^{n} \{[y_i|\theta, \sigma_E, \xi_i]\ [\xi_i|x_i]\}\ [\theta]\ [\sigma_E].$$

Hence, integrating the above expression over ξ leads to:

$$[\theta, \sigma_E|\mathbf{x}, \mathbf{y}] \propto \prod_{i=1}^{n} \left\{ \int [y_i|\theta, \sigma_E, \xi_i]\ [\xi_i|x_i] d\xi_i \right\}\ [\theta]\ [\sigma_E].$$

However, the above integrals are untractable, due to the non-linearity of g, hence we need to use some kind of numerical approximation.

In [2], a classical Monte Carlo estimation is used:

$$\int [y_i|\theta, \sigma_E, \xi_i] \ [\xi_i|x_i] d\xi_i \widehat{=} \frac{1}{L} \sum_{\ell=1}^{L} [y_i|\theta, \sigma_E, \xi_{i,\ell}],$$

with $\xi_{i,1}, \ldots, \xi_{i,L}$ independent draws from $[\xi_i|x_i]$.

More generally, we can use *quantization* techniques, which aim at approaching a continuous distribution by a discrete one, supported on a finite set of points, called nodes [7]. In the present one-dimensional case, we simply use regularly spaced points between the bounds $x_i - \epsilon$ and $x_i + \epsilon$ of ξ_i's uniform prior.

As in the previous cases, we used the Metropolis-within-Gibbs algorithm to sample from the quantized target.

5.2 Results

We ran three parallel chains with independent initial states, the first two components being independently drawn from θ's prior distribution. The number of quantization nodes was taken equal to: $L = 10$. Again, convergence tests confirmed convergence of our algorithm to the quantized posterior.

Figure 3 illustrates the posterior distributions and predictions. These can be directly compared with the results shown 2. In addition, Table 1 summarizes for all methods tested the computation times and posterior quantiles of the calibration parameters.

As expected, quantization dramatically reduces the computation time from several hours to twenty minutes. Plus, the posterior and predictive densities obtained by joint sampling and quantization are extremely similar, and observed differences may be due to the Monte Carlo error inherent to MCMC approaches.

6 Discussion

In this paper, we describe different variants of the Metropolis-Hastings algorithm that can help solve calibration problems with uncertain inputs. In particular, we introduce an original calibration strategy well-adapted to problems with few parameters and large sample sizes, which consists in pre-integrating the target posterior over a quantized version of the input distribution.

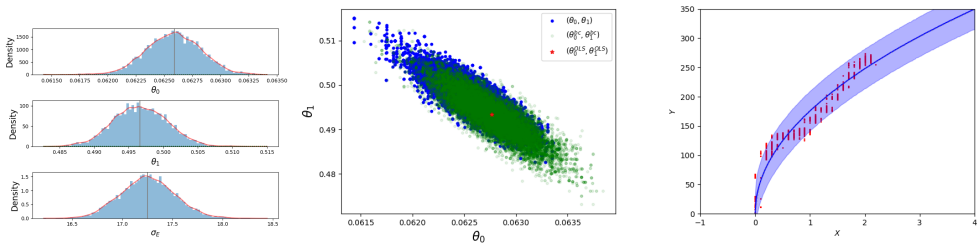


Figure 3: Results of calibration with input quantization. Left: Marginal posterior distributions of θ_0, θ_1, σ_E and ξ_1; the vertical line corresponds to the posterior median. Center: Scatter plot of the sampled θ values (blue dots), alongside the OLS estimator (red star) and the posterior density without input uncertainty (Basic Bayesian calibration, green dots). Right: Posterior 95% prediction intervals (blue) against observations (red points).

Method	time	θ_0^{LCL}	θ_0^{MAP}	θ_0^{UCL}	θ_1^{LCL}	θ_1^{MAP}	θ_1^{UCL}	σ_E^{LCL}	σ_E^{MAP}	σ_E^{UCL}
BBC	$45\,s$	6.22	6.28	6.33	4.84	4.93	5.03	20.0	20.6	21.1
joint	$5\,h\ 20\,mn$	6.22	6.27	6.31	4.88	4.95	5.03	16.8	17.4	17.9
quant	$23\,mn$	6.21	6.26	6.30	4.89	4.97	5.05	16.7	17.3	17.8

Table 1: Summary statistics for all tested algorithms: Basic Bayesian computation (BBC), joint sampling (joint), and numerical integration of inputs through quantization (quant).computation time, posterior quantiles of order 2.5% (Lower Credible Level,), 50% (Median A Posteriori) and 97.5% (Upper Credible Level). Posterior quantiles for θ_0 and θ_1 are multiplied by 100 and 10 respectively for better readibility.

In future work, we will tackle integrating the posterior density of the interest parameters over the prior density of the inputs, without updating the latter. This would require launching a potentially high number of Markov-chains in parallel, a feature yet to be developed. To limit computational load, MCMC couplings [8] could then be used to automatically reduce the chains' lengths.

Acknowledgments

The authors wish to thank their EDF colleagues Pascal Borel for providing the industrial use-case. The majority of this work has been carried out during Edimah Songo's internship at EDF R&D in Chatou in 2023.

References

[1] M. Kennedy, "Supplementary details on Bayesian Calibration of Computer Models," (2001).

[2] G. Perrin and C. Durantin, "Taking into account input uncertainties in the Bayesian calibration of time-consuming simulators," *Journal de la société française de statistique*, **160**, *2*, 24 (2019).

[3] C. P. Robert and G. Casella, *Monte Carlo statistical methods*, vol. 2, Springer (1999).

[4] M. Baudin, R. Lebrun, B. Iooss, and A.-L. Popelin, "OpenTURNS: An Industrial Software for Uncertainty Quantification in Simulation," *Handbook of Uncertainty Quantification*, 2001–2038, Springer, New York, NY.

[5] M. Keller, G. Damblin, A. Pasanisi, M. Schumann, P. Barbillon, F. Ruggeri, and E. Parent, "Validation of a Computer Code for the Energy Consumption of a Building, with Application to Optimal Electric Bill Pricing," *Econometrics*, **10**, *4* (2022); 10.3390/econometrics10040034., URL `https://www.mdpi.com/2225-1146/10/4/34`.

[6] A. Gelman and D. B. Rubin, "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, **7**, *4*, 457 (1992).

[7] C. J. Oates, "Minimum Discrepancy Methods in Uncertainty Quantification," *arXiv e-prints*, arXiv:2109.06075 (2021); 10.48550/arXiv.2109.06075.

[8] P. E. Jacob, J. O'Leary, and Y. F. Atchadé, "Unbiased markov chain monte carlo with couplings," *arXiv preprint arXiv:1708.03625* (2017).

Statistique bayésienne appliquée à la biologie

Inférence bayésienne de processus stochastiques intégrés et application à la phylogéographie virale

Paul Bastide[1], Pauline Rocu[2], Johannes Wirtz[3], Gabriel W. Hassler[4], François Chevenet[5], Denis Fargette[6], Marc A. Suchard[7], Simon Dellicour[8,9], Philippe Lemey[9] & Stéphane Guindon[2]

[1] *Université Paris Cité, CNRS, MAP5, F-75006 Paris, France,* paul.bastide@u-paris.fr
[2] *Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier. CNRS - UMR 5506. Montpellier, France,* stephane.guindon@lirmm.fr
[3] *CEFE, Université de Montpellier, CNRS, EPHE, IRD, Montpellier, France*
[4] *Department of Economics, Sociology, and Statistics, RAND, Santa Monica, CA, USA*
[5] *MIVEGEC, IRD, CNRS, Université de Montpellier, Montpellier, France*
[6] *PHIM, IRD, INRAE, CIRAD, Université de Montpellier, Montpellier, France*
[7] *Department of Biostatistics, Human Genetics and Computational Medicine, University of California, Los Angeles, CA, USA*
[8] *Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, Brussels, Belgium*
[9] *Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory for Clinical and Epidemiological Virology, KU Leuven, Leuven, Belgium*

Résumé. Au cours d'une épidémie, certains virus évoluent de manière rapide, et portent ainsi une trace dans leur génome de leurs modes de propagation. L'un des objectifs de la phylodynamique est la reconstruction d'un arbre de transmission à partir de données datées de séquences virales, notamment par le biais d'approches phylogénétiques. Lorsque les séquences sont associées à des positions géographiques, la phylogéographie permet une étude spatiale de l'épidémie, et en particulier l'inférence de sa localisation d'émergence ou d'introduction, et sa vitesse de propagation. Conditionnellement à un arbre phylogénétique daté, la phylogéographie modélise la propagation spatiale par des processus stochastiques continus en temps et en espace qui courent sur l'arbre, chaque bifurcation donnant naissance à un nouveau processus indépendant. Le mouvement brownien dit "relâché" (Lemey *et al.*, 2010), qui est très utilisé dans le domaine, est équivalent sous certaines hypothèses à un processus de saut pur. Il est ainsi, tout comme le mouvement brownien simple, à variations infinie, ce qui rend l'estimation d'une vitesse de propagation non consistante, ce qui a pu être observé dans la littérature. Dans cette communication, qui prend appui sur un travail récent (Bastide *et al.*, 2024), nous décrivons les mouvement brownien et processus d'Ornstein–Uhlenbeck intégrés, qui sont par définition plus réguliers, et qui peuvent être plus adaptés pour modéliser un mouvement spatial. Nous montrons que la vraisemblance de ces processus courant sur un arbre donné peut être calculée en temps linéaire en le nombre d'observations, ce qui rend possible leur inférence bayésienne. Nous appliquons ces résultats à l'étude de la propagation du virus du Nil Occidental en Amérique du Nord au début des années 2000.

Mots-clés. Processus stochastique, Inférence bayésienne, Phylogéographie

Abstract. Over the course of an outbreak or epidemic, many viral pathogens are known to evolve rapidly, leaving an imprint of the pattern of spread in their genomes. Uncovering the molecular footprint of this transmission process is a key goal of phylodynamic inference. Phylogeography takes dated and spatially referenced sequences to analyse the spatial spread of the epidemics, trying to answer questions such as how fast the virus is spreading, or where did it emerge or was introduced originally. Given a dated phylogenetic tree, phylogeography models the spatial spread as a continuous space and time stochastic process on the tree, that splits into independent processes at each tree bifurcation. The so-called "relaxed random walk" (Lemey *et al.*, 2010) is widely popular in the domain. It can be shown to be equivalent, under some assumptions, to a pure jump process. Just as the regular Brownian motion, it hence has infinite variations, making any speed estimation based on it inconsistent, a phenomenon that has been observed in the literature. During this talk, which relies on a recent paper (Bastide *et al.*, 2024), we introduce the integrated Brownian and Ornstein–Uhlenbeck processes, that, by definition, have more regularity, and threfore can be better suited to model a spatial spread. We show how the likelihood of such processes can be computed in a linear time in the number of observations, making Bayesian inference of such processes possible. We apply our results to the study of the spread of the West Nile Virus in North America in the early 2000s.

Keywords. Stochastic processes, Bayesian inference, Phylogeography

1 Modèles d'évolution phylogéographiques

Inférence phylogénétique bayésienne. Lors du suivi d'une épidémie, on suppose que l'on échantillonne deux types de données : les séquences moléculaires datées $\mathbf{S}$ d'une part, et un certain nombre de traits quantitatifs $\mathbf{Y}$, comme la position géographique, d'autre part. L'objectif de la phylogéographie est alors d'obtenir des informations sur l'arbre phylogénétique $\mathcal{T}$ liant les n séquences virales échantillonnées entre elles, ainsi que sur les paramètres $\boldsymbol{\theta}$ des processus d'évolution spatiaux. Dans une perspective bayésienne, on cherche à obtenir des informations sur la loi a posteriori $p(\boldsymbol{\theta}, \mathcal{T} \mid \mathbf{Y}, \mathbf{S})$. On fait ici l'hypothèse que, conditionnellement à l'arbre $\mathcal{T}$, les séquences $\mathbf{S}$ et les traits $\mathbf{Y}$ sont indépendants. Cette hypothèse, classique dans la littérature (Lemey et al., 2010), est principalement justifiée pour des raisons d'efficacité computationnelle, car elle permet de séparer l'inférence en deux problèmes distincts, en écrivant :

$$p(\boldsymbol{\theta}, \mathcal{T} \mid \mathbf{Y}, \mathbf{S}) \propto p(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T})\, p(\boldsymbol{\theta}, \mathcal{T}) = p(\mathbf{Y} \mid \boldsymbol{\theta}, \mathcal{T})\, p(\boldsymbol{\theta}) \times p(\mathbf{S} \mid \mathcal{T})\, p(\mathcal{T}) \quad (1)$$

Le deuxième terme du produit porte sur l'inférence phylogénétique classique à partir de séquences moléculaires, pour laquelle de nombreuses méthodes existent (Suchard et al., 2018). On s'intéresse ici au premier terme du produit, qui porte sur la dynamique des traits continus $\mathbf{Y}$ conditionnellement à un arbre daté $\mathcal{T}$.

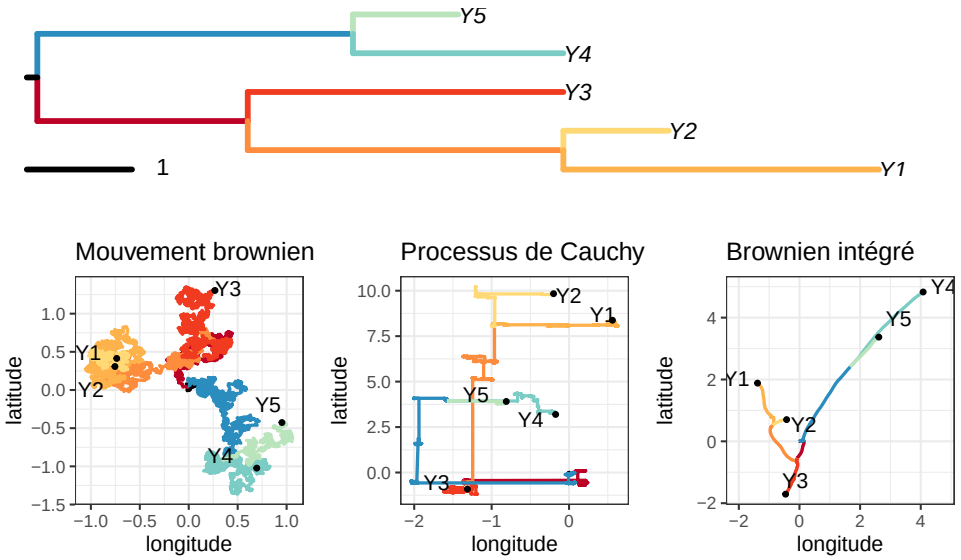


Figure 1: Réalisations d'un mouvement brownien, d'un processus de Cauchy et d'un mouvement brownien intégrés en deux dimensions (ligne du bas) sur un même arbre (ligne du haut). Les couleurs des branches de l'arbres sont reportées sur les processus.

Mouvement brownien et mouvement brownien relâché. Étant donné un arbre daté et binaire ayant n feuilles, on modélise l'évolution en temps d'un vecteur de traits $\mathbf{Y}(t)$ continus par un processus stochastique courant sur les branches de l'arbre (voir fig. 1). À un embranchement, on suppose que le processus se divise en deux trajectoires indépendantes. Les données recueillies consistent alors en la mesure de ce processus aux feuilles de l'arbre. Dans ce cadre, l'évolution suivant un mouvement brownien multivarié de dimension p (en phylogéographie, $p = 2$ pour la position spatiale) et et de variance $\mathbf{\Sigma}$ peut être défini indépendamment pour chaque branche de longueur t_i allant d'un nœud $\text{pa}(i)$ à un nœud i $(1 \leq i \leq 2n-2)$ par :

$$\mathbf{Y}_i(\tau_{\text{pa}(i)} + t) \;\big|\; \mathbf{Y}_{\text{pa}(i)}(\tau_{\text{pa}(i)}) = \mathbf{Y}_{\text{pa}(i)}(\tau_{\text{pa}(i)}) + \int_{\tau_{\text{pa}(i)}}^{\tau_{\text{pa}(i)}+t} \mathbf{\Sigma}^{1/2} d\mathbf{B}_t \;, \quad \tau_{\text{pa}(i)} \leq t \leq \tau_{\text{pa}(i)} + t_i,$$

où est temps $\tau_{\text{pa}(i)}$ écoulé entre la racine de l'arbre et $\text{pa}(i)$. Par propriété du mouvement brownien, en notant $\mathbf{Y}_i = \mathbf{Y}_i(\tau_i)$, cette caractérisation revient à supposer que, pour chaque branche de l'arbre, on a une hérédité gaussienne donnée par : $\mathbf{Y}_i \;\big|\; \mathbf{Y}_{\text{pa}(i)} \sim \mathcal{N}\big(\mathbf{Y}_{\text{pa}(i)}, t_i\mathbf{\Sigma}\big)$.

Le modèle du mouvement brownien relâché (*relaxed random walk*, Lemey et al., 2010) consiste à reprendre cette loi d'hérédité, en ajoutant, pour chaque branche, un paramètre ϕ_i qui multiplie la variance de l'incrément sur la branche. Cela ajoute de la flexibilité au modèle, qui peut "accélérer" (ϕ_i grand) ou "décélérer" sur chaque branche, et n'est plus contraint par une diffusion simple. Dans une perspective bayésienne, les ϕ_i sont vus comme des variables aléatoires, tirées indépendamment suivant une loi de régularisation. En particulier, le modèle hiérarchique suivant:

$$\phi_i \sim \text{Inv-Gamma}(1/2, t_i/2) \qquad ; \qquad \mathbf{Y}_i \;\big|\; \mathbf{Y}_{\text{pa}(i)} \sim \mathcal{N}\big(\mathbf{Y}_{\text{pa}(i)}, \phi_i t_i \mathbf{\Sigma}\big)$$

est équivalent à un processus de Cauchy, qui est un processus de sauts pur dont la loi d'hérédité est donnée par $\mathbf{Y}_i \mid \mathbf{Y}_{\mathrm{pa}(i)} \sim \text{Cauchy}\left(\mathbf{Y}_{\mathrm{pa}(i)}, t_i\boldsymbol{\Sigma}\right)$ (Bastide and Didier, 2023). Ce processus de sauts est intéressant d'un point de vue de la modélisation, car il permet de rendre compte de mouvements soudains, qui voit l'épidémie se propager d'un point à l'autre de la carte en un laps de temps très court (voir fig. 1).

Vitesse et processus intégrés. Le mouvement brownien et le processus de Cauchy sont deux processus à variation infinies, qui par définition n'admettent donc pas de vitesse. Les études utilisant le mouvement brownien relâché pour estimer des vitesses de propagation d'épidémies, bien que nombreuses, sont donc vouées à produire des estimations non consistantes, qui croissent artificiellement avec la densité d'échantillonnage (Dellicour et al., 2024). Pour pallier ce problème, on peut utiliser un processus intégré, qui modélise directement la vitesse $\mathbf{V}(t)$ comme un processus stochastique, et voit le trait mesuré $\mathbf{Y}(t) = \mathbf{Y}(0) + \int_0^t \mathbf{V}(t)dt$ comme l'intégrale en temps de cette vitesse. Ce type de processus a par exemple été utilisé en écologie, pour modéliser le mouvement d'individus sur une carte (Johnson et al., 2008). Le processus obtenu n'est pas markovien, puisque le trait mesuré dépend de la vitesse à tous les temps précédents. Cependant, le vecteur joint $\mathbf{Z}(t) = (\mathbf{V}(t); \mathbf{Y}(t))$ peut lui être markovien. Sur un arbre, en utilisant les notations précédentes, le mouvement brownien intégré peut ainsi être défini sur chaque branche par :

$$\begin{pmatrix}\mathbf{V}_i\\ \mathbf{Y}_i\end{pmatrix} \;\middle|\; \begin{pmatrix}\mathbf{V}_{\mathrm{pa}(i)}\\ \mathbf{V}_{\mathrm{pa}(i)}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mathbf{V}_{\mathrm{pa}(i)}\\ t_i\mathbf{V}_{\mathrm{pa}(i)} + \mathbf{Y}_{\mathrm{pa}(i)}\end{pmatrix}, \begin{pmatrix}t_i \cdot \boldsymbol{\Sigma} & t_i^2/2 \cdot \boldsymbol{\Sigma}\\ t_i^2/2 \cdot \boldsymbol{\Sigma} & t_i^3/3 \cdot \boldsymbol{\Sigma}\end{pmatrix}\right).$$

Comme présenté figure 1, ce processus intégré est plus lisse que le mouvement brownien, mais, de par sa variance cubique en temps, autorise des mouvements rapides sur la carte. De la même manière, le processus d'Ornstein-Uhlenbeck intégré peut être écrit sous la forme :

$$\mathbf{Z}_i \mid \mathbf{Z}_{\mathrm{pa}(i)} \sim \mathcal{N}\left(\mathbf{q}_i\mathbf{Z}_{\mathrm{pa}(i)} + \mathbf{r}_i, \boldsymbol{\Sigma}_i\right), \tag{2}$$

avec $\mathbf{q}_i$, $\boldsymbol{\Sigma}_i$ et $\mathbf{r}_i$ deux matrices et un vecteur qui ne dépendent que des paramètres du processus et de la longueur de la branche (Bastide et al., 2024).

2 Inférence et application

Calcul de la vraisemblance. Ces deux processus étant gaussiens, on peut écrire de manière explicite la loi jointe de tous les traits mesurés aux feuilles, qui suit une loi normale multivariée, et ainsi obtenir la vraisemblance $p\left(\mathbf{Y} \mid \boldsymbol{\theta}, \mathcal{T}\right)$ qui intervient dans (1). Les formules sont données dans Bastide et al. (2024), et impliquent l'inversion de matrices de taille $n \times n$. Pour rendre les calculs plus efficaces, on peut se baser sur l'équation de propagation (2), qui est linéaire gaussienne. Comme démontré par Mitov et al. (2020), un algorithme récursif de type filtre de Kalman permet dans ce cas de calculer la vraisemblance en un temps linéaire en n. Dans ce même cadre, Bastide et al. (2021) montrent qu'un second parcours de l'arbre permet d'obtenir la loi des états ancestraux non observés conditionnellement aux feuilles. De manière cruciale, cet algorithme, implémenté dans le logiciel BEAST (Suchard

et al., 2018), permet de gérer les données manquantes. Dans le cas des processus intégrés (2), seul le trait $\mathbf{Y}$ est observé, si bien que chaque vecteur $\mathbf{Z}$ aux feuilles contient des composantes manquantes (la moitié est non observée). Les deux parties du vecteur étant corrélées, il reste possible d'obtenir des informations sur la loi des vitesses non observées conditionnellement aux positions observées, ce qui permet de reconstruire la dynamique spatiale de l'épidémie.

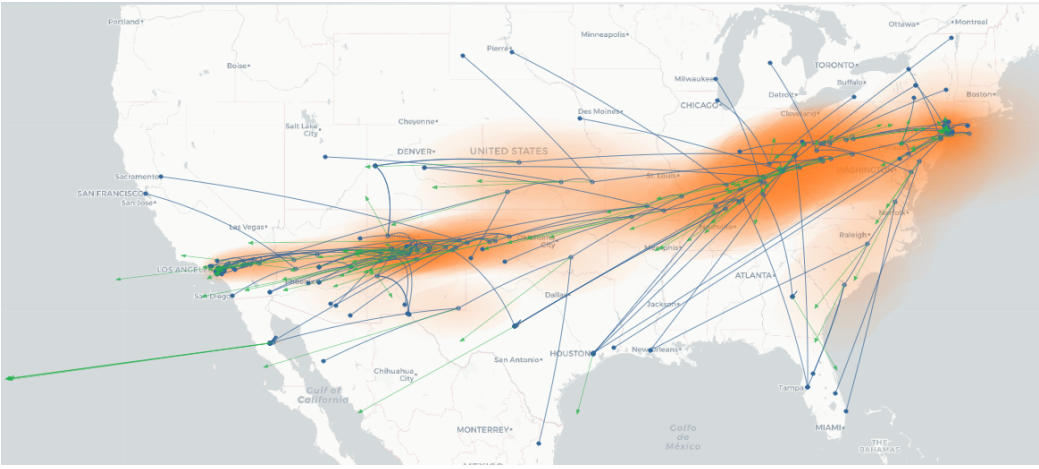


Figure 2: Reconstruction de la dynamique spatiale de l'épidémie du virus du Nil occidental en Amérique du Nord au début des années 2000. Les points bleus pleins représentent les positions des séquences échantillonnées, et les points bleus vides les positions ancestrales reconstruites, entourés par une zone de crédibilité à 80% en orange. Les lignes bleues relient chaque nœud à son parent dans l'arbre phylogénétique reconstruit. Les flèches vertes représentent les vitesses ancestrales estimées. La figure est tracée par Evolaps v2.42 (Chevenet et al., 2024).

Application au virus du Nil occidental. L'épidémie de virus du Nil occidental, partie de la région de New York en 1999, a fait plus d'un millier de morts aux États-Unis et a été très suivie par les autorités sanitaires. On reprend ici le jeu de données analysé par Pybus et al. (2012), qui contient 104 séquences datées et localisées. On utilise le même modèle d'évolution de séquence et d'horloge, et l'on remplace le modèle géographique par un mouvement brownien intégré. L'inférence bayésienne est faite à l'aide de BEAST (Suchard et al., 2018). Une représentation du processus reconstruit sur l'arbre des clades de crédibilité maximale (*MCC tree*) est présentée figure 2.

Perspectives. Nous présentons ici et dans Bastide et al. (2024) une méthode efficace pour l'inférence de processus intégrés en phylogéographie. D'un point de vue appliqué, il serait intéressant de voir dans quels cas ces processus intégrés sont plus adaptés que les processus browniens relâchés actuellement utilisés, et si les vitesses estimées peuvent être utilisées pour réaliser des prédictions. Du point de vue de la modélisation, l'influence de variables spatio-temporelles ne peut être traitée dans tous les modèles présentés ici que de manière post-hoc (Dellicour et al., 2020). L'intégration directe de covariables dans le modèle pourrait permettre de tester des hypothèses biologiques de manière plus fine.

Bibliographie

Bastide, P. and Didier, G. (2023). The cauchy process on phylogenies: A tractable model for pulsed evolution. *Systematic Biology*, 72(6):1296–1315.

Bastide, P., Ho, L. S. T., Baele, G., Lemey, P., and Suchard, M. A. (2021). Efficient Bayesian Inference of General Gaussian Models on Large Phylogenetic Trees. *The Annals of Applied Statistics*, 15(2):971–997.

Bastide, P., Rocu, P., Wirtz, J., Hassler, G. W., Chevenet, F., Fargette, D., Suchard, M. A., Dellicour, S., Lemey, P., and Guindon, S. (2024). Modeling the velocity of evolving lineages and predicting dispersal patterns. *Proceedings of the National Academy of Sciences*, 121(47):2024.06.06.597755.

Chevenet, F., Fargette, D., Bastide, P., Vitré, T., and Guindon, S. (2024). Evolaps 2: Advanced phylogeographic visualization. *Virus Evolution*, 10(1).

Dellicour, S., Bastide, P., Rocu, P., Fargette, D., Hardy, O. J., Suchard, M. A., Guindon, S., and Lemey, P. (2024). How fast are viruses spreading in the wild? *PLOS Biology*, 22(12):e3002914.

Dellicour, S., Lequime, S., Vrancken, B., Gill, M. S., Bastide, P., Gangavarapu, K., Matteson, N. L., Tan, Y., du Plessis, L., Fisher, A. A., Nelson, M. I., Gilbert, M., Suchard, M. A., Andersen, K. G., Grubaugh, N. D., Pybus, O. G., and Lemey, P. (2020). Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. *Nature Communications*, 11(1).

Johnson, D. S., London, J. M., Lea, M.-A., and Durban, J. W. (2008). Continuous-Time Correlated Random Walk Model for Animal Telemetry Data. *Ecology*, 89(5):1208–1215.

Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, 27(8):1877–1885.

Mitov, V., Bartoszek, K., Asimomitis, G., and Stadler, T. (2020). Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts. *Theoretical Population Biology*, 131:66–78.

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., Gray, R. R., Arinaminpathy, N., Stramer, S. L., Busch, M. P., and Delwart, E. L. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences*, 109(37):15066–15071.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1):1–5.

Modélisation de l'évolution de la présence d'une infection par un modèle bayésien de Markov caché à observations continues. Application à la surveillance de la BVD en France

Aurélien Madouasse[1] & Matthieu Trotreau & Grégoire Kuntz[2] & Jean-Michel Galharret[3]

[1] *Oniris, INRAE, BIOEPAR, Nantes 44300, France, aurelien.madouasse@oniris-nantes.fr*
[2] *Innoval, GDS Bretagne, gregoire.kuntz@innoval.com*
[3] *Oniris, INRAE, StatSC, Nantes 44300, France, jean-michel.galharret@oniris-nantes.fr*

Résumé. Un modèle bayésien de Markov caché a été développé pour estimer les caractéristiques de tests sérologiques et la dynamique d'une infection à partir de données longitudinales de surveillance. L'approche proposée consiste à modéliser les résultats de tests comme un mélange de deux distributions (l'une pour les échantillons séronégatifs et l'autre pour les séropositifs). En l'absence de seuil de détection, une règle de décision basée sur les paramètres du mélange est proposée. Cette approche est comparée par simulations à l'approche classique (binarisation du résultat du test comme observation de la variable latente). Les résultats montrent une bonne estimation des paramètres du mélange et de meilleures performances que le modèle classique dans tous les scénarios épidémiologiques considérés. Le modèle a été appliqué aux données du programme de surveillance du virus de la diarrhée virale bovine (BVD) en Bretagne à partir de résultats trimestriels de deux tests sérologiques sur le lait de tank entre 2014 et 2020.

Mots-clés. Modèle bayésien de Markov caché, mélange de lois.

Abstract. A Bayesian Hidden Markov Model (BHMM) has been developed to estimate test characteristics and infection dynamics from longitudinal surveillance data. The proposed approach involves modeling test results as a mixture of two distributions (one for seronegative samples and the other for seropositive). In the absence of a detection threshold, we propose a decision rule based on the parameters of this mixture. This approach is compared through simulations to the classical approach (binarization of the test result as an observation of the latent variable). The results show an accurate estimation of the mixture parameters and outperforms the classical model across all considered epidemiological scenarios. The model was applied to data from the Bovine Viral Diarrhea (BVD) virus surveillance program in Brittany, using quarterly ELISA test results from bulk tank milk between 2014 and 2020.

Keywords. Hidden Markov Model, Gaussian mixture ...

1 Modèle Gaussien de Markov caché

1.1 Introduction

Le modèle de Markov caché (Hidden Markov Model, HMM), introduit par Baum (1966), est un processus stochastique utilisé pour représenter une variable catégorielle latente qui évolue en temps discret ($t = 1, \ldots, T$) selon une dynamique markovienne : l'état suivant de la chaîne ne dépend que de son état actuel et est indépendant des états précédents. À tout moment, l'état latent Z_t détermine la distribution d'une variable observée Y_t. Dans les programmes de surveillance de maladie, un HMM à deux états latents est considéré et Y_t peut être un résultat de test binaire (par exemple, positif ou négatif) ou bien le résultat d'une mesure de concentration d'anticorps (e.g. tests ELISA). Dans le premier cas, Y_t suit une distribution de Bernoulli avec un paramètre p dépendant de Z_t (see Madouasse et al., 2022). Dans le second cas, Y_t peut être modélisée par une distribution de probabilité.

Un HMM a deux états latents est caractérisé par la prévalence initiale de la maladie $\pi_1 = P(Z_1 = 1)$ et deux probabilités de transition $\tau_1 = P(Z_t = 1|Z_{t-1} = 0)$, $\tau_2 = P(Z_t = 1|Z_{t-1} = 1)$. Ces probabilités sont supposées être indépendantes du temps (i.e., les chaines sont homogènes). On suppose que les résultats de tests sont distribués selon un mélange gaussien de paramètres (μ_0, σ_0) (séronégatifs) et (μ_1, σ_1) (séropositifs).

$$P(Y < y|Z = z) = (1 - z)\Phi_0(y) + z\Phi_1(y),$$

où $z = 0, 1$, $y \in \mathbb{R}$, et Φ_i est la fonction de répartition de la loi normale de paramètres (μ_i, σ_i). La figure 1 résume ce modèle gaussien de Markov caché.

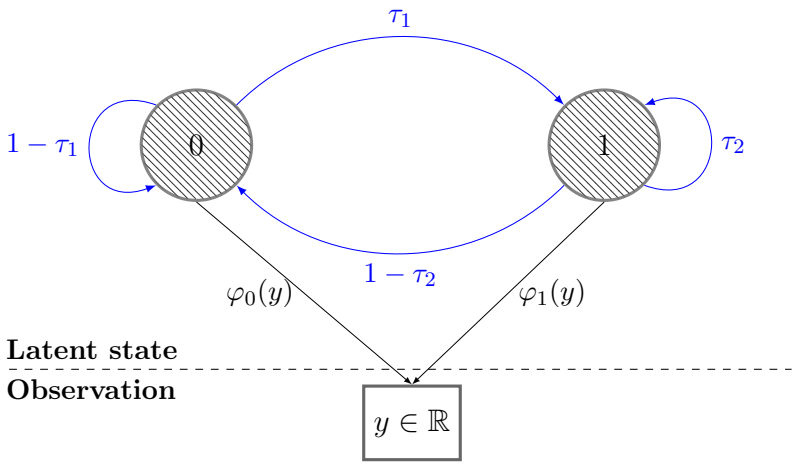


FIGURE 1 – HMM à deux états avec des données observées distribuées selon un mélange gaussien. φ_i correspond à la densité gaussienne de moyenne μ_i et de variance σ_i^2.

1.2 Inférence bayésienne

Soit $\theta = (\pi_1, \tau_1, \tau_2, \mu_0, \sigma_0, \mu_1, \sigma_1)$ l'ensemble des paramètres du modèle. Étant donné une séquence de T observations $Y_{1:T} = Y_1, \ldots, Y_T$, six types d'inférence sont effectués :

Filtering : Calculer pour $t < T$, la probabilité de l'état Z_t conditionnement aux observations $Y_{1:t}$.

Smoothing : Calculer la probabilité de Z_t conditionnellement à toutes les observations $Y_{1:T}$.

Forecasting : Pour $h \in \mathbb{N}$, calculer la probabilité de Z_{t+h} conditionnellement aux observations $Y_{1:t}$.

Evaluation : Calculer la vraisemblance de la séquence $Y_{1:T}$.

Decoding : Déterminer la séquence $Z_{1:T}$ la plus probable étant données les observations $Y_{1:T}$.

Learning : Estimer le paramètre θ sachant $Y_{1:T}$.

L'algorithme forward, l'algorithme backward et l'algorithme de Viterbi sont utilisés pour résoudre ces problèmes (voir par exemple Viterbi (1967); Rabiner (1989)). Parmi ces tâches, l'estimation de θ (*Learning*) peut être réalisée par maximum de vraisemblance, l'algorithme EM ou l'inférence bayésienne (voir par exemple Ghahramani, 2001).

La distribution jointe des observations $Y_{1:T}$ et des états latents $Z_{1:T}$ conditionnement à θ est

$$P(Y_{1:T}, Z_{1:T}|\theta) = P(Z_1)P(Y_1|Z_1, \mu_0, \sigma_0, \mu_1, \sigma_1) \prod_{t=2}^{T} P(Z_t|Z_{t-1})P(Y_t|Z_t, \mu_0, \sigma_0, \mu_1, \sigma_1),$$

où $P(Z_1)$ dépend uniquement de π_1, $P(Z_t|Z_{t-1})$ uniquement de τ_1, τ_2. De plus on a : $P(Y_{1:T}|\theta) = \sum_{Z_{1:T}} P(Z_{1:T}Y_{1:T}|\theta)$ dont on déduit la loi a posteriori de $\theta|Y_{1:T}$.

1.3 Règle de décision basée sur la donnée continue

Soit Z l'état latent correspondant à un résultat de test y. On adopte la règle de décision suivante :

$$Z = \begin{cases} 1 \text{ lorsque } \Phi_0(y) > 1 - \Phi_1(y), \\ 0 \text{ sinon.} \end{cases} \tag{1}$$

Un calcul simple montre que si $s = \frac{\sigma_0\mu_1 + \sigma_1\mu_0}{\sigma_0 + \sigma_1}$, alors on a $\Phi_0(s) = 1 - \Phi_1(s)$. Donc cette règle de décision est équivalente à déclarer un résultat de test y positif lorsque $y > s$. Par ailleurs, la sensibilité (i.e. le taux de vrais positifs) et la spécificité (i.e. le taux de faux négatifs) associées à ce seuil sont égales à $\Phi_0(s) = 1 - \Phi_1(s)$. La loi a posteriori de s peut être déduite de celles des paramètres du mélange et servir de référence en l'absence de seuil de détection.

Scenario	τ_1	τ_2	μ_0	μ_1	Seuil (h)	Se	Sp
1	0.1	0.85	25	75	50	0.95	0.95
2	0.1	0.85	25	50	37.5	0.79	0.79
3	0.1	0.85	30	50	37.5	0.79	0.69
4	0.1	0.85	25	45	37.5	0.69	0.79
5	0.2	0.7	25	75	50	0.95	0.95
6	0.4	0.4	25	75	50	0.95	0.95

TABLE 1 – Valeurs des paramètres utilisés pour les simulations.

2 Résultats

2.1 Données simulées

Pour les simulations les écarts types des deux distributions du mélange ont été fixés à $\sigma_0 = \sigma_1 = 15$. La prévalence initiale de la maladie a été fixée à $\pi_1 = 0.4$. La table 1 fournit les valeurs des autres paramètres utilisés pour les simulations. Le choix de τ_2 a été réalisé tel que $\tau_2 = 1 - \frac{\tau_1(1-\pi_1)}{\pi_1}$ qui est standard en épidémiologie car il correspond à une situation où l'infection est équilibrée (i.e. le nombre d'individus éliminant l'infection est équilibré par le nombre d'individus l'acquérant). La valeur du seuil h permet de transformer les données en résultats binaires (i.e. $Z = 0$ pour $y < h$) et ainsi utiliser le modèle proposé par Madouasse et al. (2022). La sensibilité et la spécificité de ce seuil h sont respectivement égales à $1 - \Phi_1(h)$ et $\Phi_0(h)$. Les scénarios 1 à 4 correspondent à un chevauchement de plus en plus important entre les deux composantes du mélange. Les scénarios 5 et 6 correspondent à une augmentation de la probabilité de devenir infecté (respectivement $\tau_1 = 0.2$ et $\tau_1 = 0.4$). Le graphe 2 fournit les résultats de la comparaison des deux modèles sur $B = 1500$ échantillons de $n = 100$ individus suivis sur $t = 5$ périodes. Les performances des modèles sont comparables en terme de biais sur les quatre premiers scénarios. Par contre le biais des estimations avec le modèle à réponses binaires est très mauvais pour les paramètres τ_1, τ_2 dans les scénarios 5 et 6 (haut du graphe), alors que celui des estimations obtenues avec le modèle à observations continues reste équivalent aux autres scénarios (1 à 4). Par ailleurs, les écarts types des distributions a posteriori des paramètres du modèle à données continues sont dans tous les scénarios inférieurs à ceux du modèle à données discrètes (bas du graphe).

2.2 Données réelles

Les données sont issues d'un plan de surveillance de la BVD (Diarrhée Bovine Virale) réalisé sur des troupeaux entre 2014 et 2020 dans la région Bretagne à l'aide de deux tests de détection d'anticorps appelés BTMab1 et BTMab2. Les troupeaux ont été testés tous les trois mois, les deux tests étant utilisés en alternance. Étant donné que les tests mesurent la présence d'anticorps dans le tank de lait (sample-pooling), l'état de latence considéré dans cette étude a été défini comme *la présence d'au moins un animal séropositif dans le troupeau*. Le seuil utilisé pour les deux tests avait été fixé à 35 pour toute la durée de l'étude. Les paramètres du modèle ont été estimés chaque année (4 temps de mesure), l'année $(t - 1)$

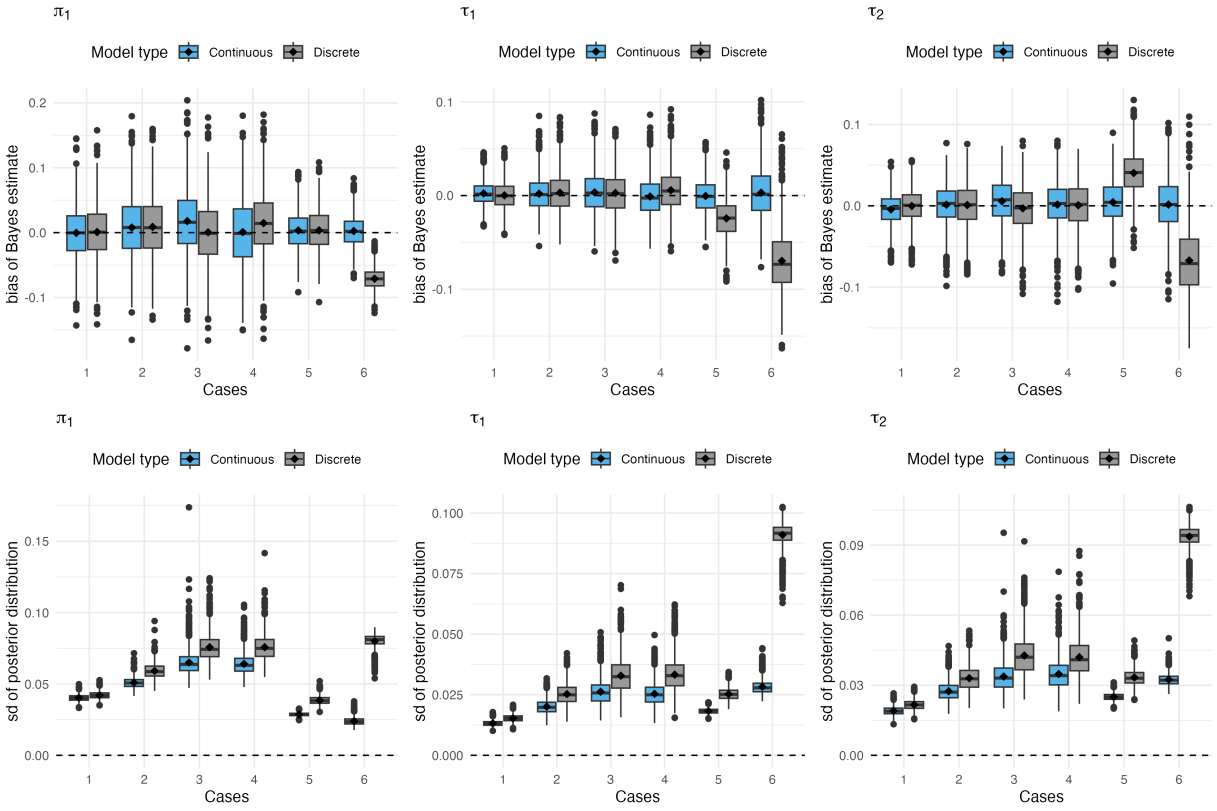


FIGURE 2 – Boxplots du biais de l'estimateur de Bayes des paramètres de dynamique π_1, τ_1, τ_2 (haut) et des écarts types de leur distribution a posteriori (bas) en fonction du scénario considéré et du modèle Continuous : modèle à observations continues) et Discrete : modèle à observations binaires).

servant d'historique pour l'année t. La figure 3 fournit les estimations des caractéristiques des deux tests durant la période d'étude. On constate que la moyenne de la composante des troupeaux infectés du test BTMab1 est proche voir en dessous du seuil de détection dans la plupart des départements. Ainsi, le fait d'utiliser un seuil constant alors qu'il s'agit de données de pooling et qu'on ne tient donc en compte de la prévalence intra-troupeaux de l'infection va détériorer les performances des tests en terme de sensibilité. En revanche, utiliser la règle de décision (1) associée aux données continues aurait permis de garantir une sensibilité minimale donnée par la table 2.

Références

Ghahramani, Z. (2001). An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(01) :9–42.

Madouasse, A., Mercat, M., van Roon, A., Graham, D., Guelbenzu, M., Santman Berends, I., van Schaik, G., Nielen, M., Frossling, J., Agren, E., Humphry, R., Eze, J., Gunn, G.,

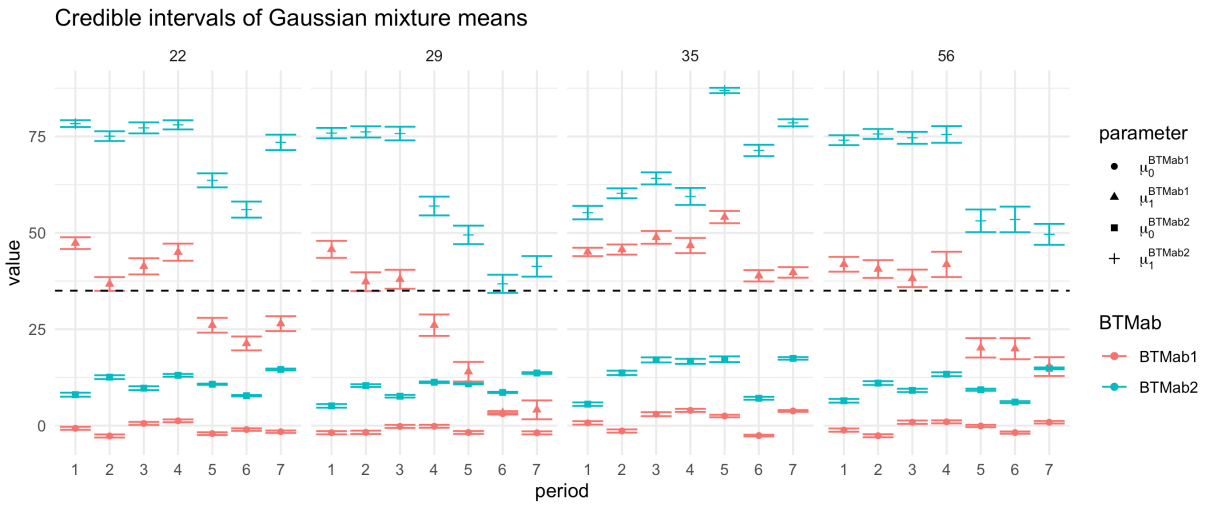


FIGURE 3 – Evolution des paramètres des tests sérologiques au cours de la période d'étude en fonction du département breton.

	période	Seuil	$S_e = S_p$
1	1	17.01	0.90
2	2	10.78	0.85
3	3	13.93	0.85
4	4	12.54	0.85
5	5	4.22	0.75
6	6	3.54	0.71
7	7	4.98	0.76

TABLE 2 – Evolution du seuil de décision de la règle (1) pour le département 22 entre 2014 et 2020. S_e et S_p désignent respectivement la sensibilité et la spécificité du seuil issue de la règle de décision associée.

Henry, M. K., Gethmann, J., More, S. J., Toft, N., and Fourichon, C. (2022). A modelling framework for the prediction of the herd-level probability of infection from longitudinal data. *Peer Community Journal*, 2.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2) :260–269.

Bayesian estimation of brain functional connectivity networks

Yiye Jiang [1], Alice Chevaux [1], Julyan Arbel[1], Guillaume Kon Kam King[2], Wendy Meiring[3], Alex Petersen[4], Sophie Achard[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, yiye.jiang@inria.fr, alice.chevaux@inria.fr, julyan.arbel@inria.fr, sophie.achard@cnrs.fr*
[2] *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France, guillaume.konkamking@inrae.fr*
[3] *University of California Santa Barbara, Santa Barbara, 93106, CA, USA, meiring@ucsb.edu*
[4] *Brigham Young University, Provo, 84602, UT, USA, petersen@stat.byu.edu*

Résumé. L'inférence des réseaux de connectivité fonctionnelle (FC) cérébrale à partir des données d'IRMf de repos (rs-fMRI) est un sujet clé en neuroimagerie. Cet article présente un modèle bayésien permettant d'inférer un graphe de FC à partir de séries temporelles rs-fMRI multivariées d'un sujet. Contrairement aux méthodes existantes qui estiment des poids constants, notre modèle infère des distributions a posteriori des poids, permettant des estimations plus robustes en prenant en compte l'incertitude d'estimation et soutenant diverses analyses post-inférence. À notre connaissance, il existe un seul modèle bayésien de FC utilisable pour les données rs-fMRI d'un sujet unique, ce qui fait de notre approche une contribution précieuse. Nos expérimentations numériques démontrent des performances supérieures de notre nouvelle approche. En pratique, nous proposons une distribution a priori sur les matrices de corrélation, permettant un contrôle explicite des moyennes et des variances marginales des entrées individuelles, facilitant ainsi l'intégration des connaissances expertes.

Mots-clés. IRM fonctionnelle de repos, inférence de la connectivité fonctionnelle cérébrale, modélisation bayésienne, détection d'arêtes, priors de corrélation.

Abstract. Inference of brain functional connectivity (FC) networks from resting-state fMRI (rs-fMRI) data is a key focus in neuroimaging. This paper introduces a Bayesian model to infer a FC graph from multivariate rs-fMRI time series from one subject. Unlike most existing methods for this problem that estimate constant weights, our model infers posterior distributions of weights, yielding more robust point estimates by taking into account estimation uncertainty and enabling diverse post-inference analyses. To the best of our knowledge, only one existing Bayesian FC model is applicable to resting-state fMRI data for single subjects, making our approach a valuable addition to the field, as demonstrated by its superior performance in our experiments. In practice, we introduce a prior distribution over correlation matrices that allows explicit control over the marginal means and variances of individual entries. This flexibility not only facilitates the integration of expert knowledge but also provides a robust foundation for constructing other models in the field.

Keywords. Resting-state functional MRI, brain functional connectivity inference, Bayesian modeling, edge detection, correlation priors.

1 Introduction

Functional connectivity (FC) looks at how different parts of the brain show similar temporal characteristics. It is often measured by the level of temporal dependence during functional magnetic resonance imaging (fMRI). Statistical methods play a crucial role in learning and analyzing FC from fMRI data (Lazar 2008). Classical methods include calculating temporal correlations between regions of interest, clustering, and Gaussian Graphical Lasso. Among others, Bayesian approaches have become an emerging framework for FC inference since one of its first works (Patel et al. 2006). Advantages of the posterior distributions include providing an estimate of uncertainty, which may be used to derive more robust decision rules, and the flexibility to develop realistic models. Additionally, deterministic FC graphs do not allow as wide a variety of post-inference analyses as distributional connectivity graphs. For example, there is no robust way to compare two graphs statistically to identify the significantly different edges when the edges are constant. In frequentist FC literature, such comparison exists only for two groups of graphs, for example, in Ginestet et al. (2017). Another example is the detection of zeros in the connectivity matrices. A popular method is to detect "significant" edges by applying some multiple testing procedures (Drton & Perlman 2007). Using posterior distributions from Bayesian inference will provide a simple theoretical framework which will potentially lead to more robust results.

Although many works have shown the success of Bayesian methods in inferring FC networks, most of them are proposed for task-based fMRI. In this work, we propose a Bayesian approach to the inference of static FC networks from resting-state (rs)-fMRI data for a single subject. To our knowledge, there is only one existing approach (Acharyya et al. 2023), which we used as a comparison. The results show that the proposed method outperforms the existing approach.

The technical contribution lies in the development of a family of priors for correlation matrices, which allows for control over the mean and variance of each correlation matrix entry in an explicit way. This control facilitates the incorporation of expert knowledge and has been studied in Bayesian FC inference models for a group of subjects (Gorbach et al. 2020). However, these approaches rely on individually controlling each entry, in other words, imposing an independently univariate distribution on each entry. This way does not respect the positive-definiteness constraint of the correlation matrix. At this point, the proposed prior can realize the marginal control while taking into account the positive definiteness. Since the correlation matrix is one of the most used metrics of FC connectivity, our framework will be useful for many real case studies, which can be easily used to build other Bayesian models (such as group models) to perform the explicit exponent control. We summarize the proposed construction and the main theoretical results in the next section.

2 Single-subject functional connectivity models

Let us define wavelet-decomposed fMRI measurements at time t and region k as the random variable $\boldsymbol{x}_{kt}, t = 1, \ldots, T, k = 1, \ldots, K$ modeling wavelet coefficients. We assume a Gaussian

likelihood for $\boldsymbol{x}_{kt}$ as it is a standard assumption in the literature:

$$\boldsymbol{x}_t = \{\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{Kt}\}^\top | \boldsymbol{\Sigma} \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma}). \tag{1}$$

The iid is a convenient hypothesis which was proved to be approximately valid by performing basic wavelet routines using existing packages such as `waveslim` in R (Achard & Gannaz 2016). In addition, there is no loss of generality in assuming the mean equal to zero.

We consider the induced correlation matrix defined by $\boldsymbol{R} = \boldsymbol{\Delta}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1/2}$, where $\boldsymbol{\Delta}$ is a diagonal matrix consisting of the diagonal elements of $\boldsymbol{\Sigma}$. Let us denote by $\boldsymbol{\rho}_{kk'} = \text{Corr}(\boldsymbol{x}_{tk}, \boldsymbol{x}_{tk'})$ its entries. The variable of interest is $\boldsymbol{R}$, but we first consider a prior on $\boldsymbol{\Sigma}$ which is conjugate to the normal likelihood. We then study the properties of the induced prior on $\boldsymbol{R}$. In this context, we need a prior on the covariance $\boldsymbol{\Sigma}$, parametrized in such a way that the parameters of the prior of $\boldsymbol{\Sigma}$ can control directly and explicitly the induced prior on $\boldsymbol{R}$ and the marginal priors of $\boldsymbol{\rho}_{kk'}$. Thus, the prior we propose induces an implicit prior on $\boldsymbol{R}$. The FC application requires the following properties for the prior of $\boldsymbol{\rho}_{kk'}$:

1. Marginals of $\boldsymbol{\rho}_{kk'}$ are equal, for $k \neq k'$;
2. $\mathbb{P}(\boldsymbol{\rho}_{kk'} > 0) > \mathbb{P}(\boldsymbol{\rho}_{kk'} < 0)$;
3. Explicit control over the mean and variance of $\boldsymbol{\rho}_{kk'}$;
4. Maintaining the positive-definiteness of $\boldsymbol{R}$.

Point 2 corresponds to application-specific prior information. Indeed, when using correlation as FC metric, generally more positive values are observed.

We propose an inverse-Wishart prior $\boldsymbol{\Sigma} \sim \mathcal{W}^{-1}(\Psi, \nu)$, with degree of freedom $\nu > K + 1$ and the following structure for the hyperprior scale parameter Ψ:

$$\Psi = (\nu - K - 1)\text{diag}(\boldsymbol{\sigma})\tilde{\boldsymbol{\rho}}_0\text{diag}(\boldsymbol{\sigma}), \tag{2}$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_K)$ and $\tilde{\boldsymbol{\rho}}_0$ is the $K \times K$ matrix with diagonal elements equal to 1, and off-diagonal elements equal to $\rho_0 \in (-1/(K-1), 1)$. This reduces the number of effective parameters, corresponding to Point 1 above. Thus the inverse-Wishart prior is parametrized by ρ_0, ν, σ_k. The following theorem shows how ρ_0 and ν control $\mathbb{E}(\boldsymbol{\rho}_{kk'})$ and $\mathbb{V}(\boldsymbol{\rho}_{kk'})$. We also show that our construction provides the other desired properties listed above.

Theorem 1. *Prior (2) entails the following properties on the induced correlation matrix:*

1. $\boldsymbol{\rho}_{kk'} \overset{D}{=} \boldsymbol{\rho}_{ll'}$, *for* $k \neq k', l \neq l'$. *Furthermore,* $(\boldsymbol{\rho}_{kk'})_{k<k'}$ *are exchangeable.*
2. *When* $\rho_0 \geq 0$, $\mathbb{P}(\boldsymbol{R}_{kk'} > 0) \geq \mathbb{P}(\boldsymbol{R}_{kk'} < 0)$, *with the equality obtained by* $\rho_0 = 0$;
3. $\boldsymbol{\rho}_{kk'} \to \rho_0$ *in probability, for* $k \neq k'$ *and as* $\nu - K \to \infty$;
4. $\mathbb{E}(\boldsymbol{\rho}_{kk'}) = \rho_0\left(1 - \frac{1-\rho_0^2}{2(\nu-(K-2))}\right) + O\left(\frac{1}{(\nu-K)^2}\right)$, *for* $k \neq k'$;
5. $\mathbb{V}(\boldsymbol{\rho}_{kk'}) \approx \frac{(1-\rho_0^2)^2}{(\nu-(K-2))-1}$, *for* $k \neq k'$ *and as* $\nu - K \to \infty$.

The marginal mean is approximately equal to ρ_0, with better precision when increasing the sample size. The parameter ν plays a dual role: on the one hand, it controls the variance and greater ν leads to a smaller variance, akin to a precision parameter, on the other hand, it controls the mean as well, greater ν brings the mean closer to ρ_0. Thus, expert knowledge can be (initial belief about ρ and degree of confidence) translated into a choice of ρ_0 and ν. For the hyperparameters σ_k, without further information, we can assign the sample standard deviation, because $\mathbb{E}_{prior}(\boldsymbol{\Sigma}_{kk}) = \sigma_k^2$. In addition, whenever $\rho_0 > 0$, the prior favors positive correlations: $\mathbb{P}(\boldsymbol{\rho}_{kk'} > 0) > \mathbb{P}(\boldsymbol{\rho}_{kk'} < 0)$. When $\rho_0 = 0$, $\mathbb{P}(\boldsymbol{\rho}_{kk'} > 0) = \mathbb{P}(\boldsymbol{\rho}_{kk'} < 0)$.

The inverse-Wishart distribution is a conjugate prior for the Gaussian likelihood, so the posterior on $\boldsymbol{\Sigma}$ is explicit, which allows us to state the following result.

Theorem 2. *The inverse-Wishart prior entails the following control on the posterior marginals of the correlation matrix, for $k \neq k'$ and as $\nu - K \to \infty$:*

$$\mathbb{E}(\boldsymbol{\rho}_{kk'} \mid \text{data}) = \rho_T^{kk'} \left(1 - \frac{1 - (\rho_T^{kk'})^2}{2(\nu_T - (K-2))}\right) + O\left(\frac{1}{(\nu_T - K)^2}\right),$$
$$\mathbb{V}(\boldsymbol{\rho}_{kk'} \mid \text{data}) \approx \frac{(1 - (\rho_T^{kk'})^2)^2}{\nu_T - K + 1},$$

where

$$\nu_T = \nu + T, \quad \rho_T^{kk'} = \frac{\Psi_T^{kk'}}{\left(\Psi_T^{kk}\Psi_T^{k'k'}\right)^{1/2}}, \quad \textit{with} \;\; \Psi_T = \Psi + \sum_{t=1}^{T} \boldsymbol{x}_t^\top \boldsymbol{x}_t.$$

The proofs of Theorem 1 and Theorem 2 are provided in Jiang et al. (2025).

3 Application on the rats data set

In this section, we apply our method on a real data set to illustrate the effect of the prior parameters ρ_0 and ν. We fix σ_k by sample standard deviation. Then we highlight the usefulness of posterior distributions by showing an application to edge detection. We also compare with the competitor Acharyya et al. (2023).

Rats data set. The data concern experiments to evaluate the influence of anesthetics on FC for brain rats. Time series of 3600 time points are extracted from rs-fMRI on brain areas defined in an atlas, consisting of 51 regions. After the preprocessing using wavelets, we have 219 time steps. In this work, we show the results for a dead rat and a live rat.

Figure 1 reports the posterior modes of correlation of the two rats under different ρ_0 and ν values. When ν is set to a small value, corresponding to light prior, ρ_0 has little effect on the posterior distribution. By contrast with a stronger level of prior, $\rho_0 = 0.3$ shifted the posterior modes to more positive values as shown in the two last subfigures. We see clear patterns for the FCs of live rat in comparison of those of the dead rat which are almost noise. In addition, the patterns change across the subfigures which implies that the prior impact on the posterior distribution is not linear.

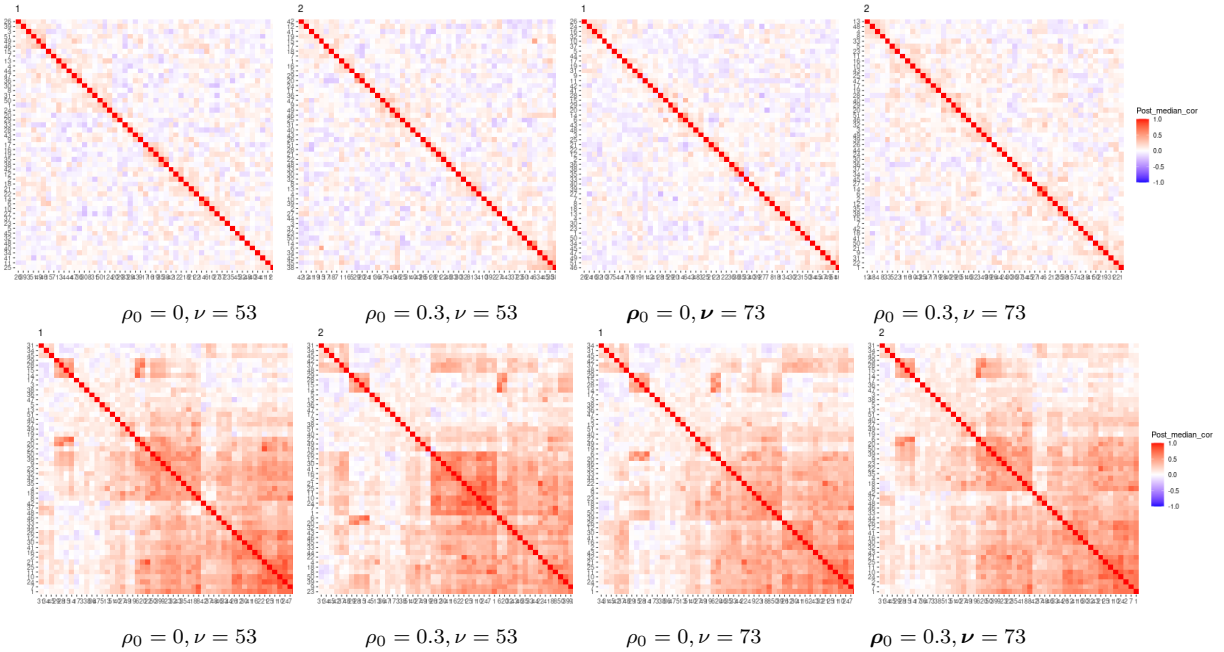
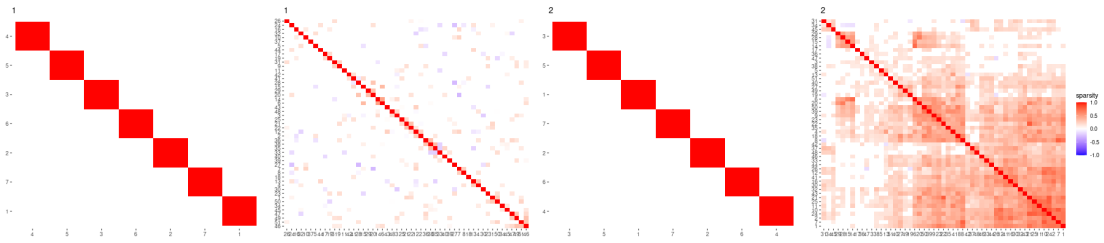


Figure 1: Posterior mode correlation matrices (MAP estimation) from the two rats (size = 51). Top: dead rat. Below: live rat. The reordering of regions comes from the hierarchical clustering of row (thus column) vectors.

When no expert knowledge is available, we can use information criteria to select the hyperparameters ρ_0, ν. We use Bayesian leave-one-out estimate of out-of-sample predictive fit proposed in Vehtari et al. (2017). The resulting best fits, respectively for the dead and live rats, are reported in Figure 1 with bold font. We can see that for the dead rat, the best fit comes from the stronger belief of zero correlation ($\rho_0 = 0, \nu = 73$). For the live rat, the best fit comes from the stronger belief of positive correlation ($\rho_0 = 0.3, \nu = 73$). We proceed with the two best models.

The previous results illustrate how Bayesian inference can incorporate the expert knowledge when available. However the induced point estimation cannot fully show the advantages of posterior distributions, including uncertainty evaluation and richer post-inference analysis. In Figure 2, we show one application of the posteriors in "significant" edge detection. We construct the shortest credible set of probability 0.9 from the posterior distribution of each edge/correlation $\boldsymbol{\rho}_{kk'}$, then check whether 0 falls within the interval. If not, the edge is deemed "significant". Thus the detection takes into account the uncertainty that is variance, which gives a more robust result than thresholding point estimates. In the same figure, we also show the detection results from the competitor Acharyya et al. (2023). Since their observations consist of covariance matrices, we use the posterior predictive distributions of the induced correlation matrices. The covariance samples are calculated using the raw time series. Especially, since their model is only able to deal with small dimensions in real data context, we reduced the problem dimension to the one considered in Acharyya et al. (2023), which consists of 7 regions. We use the same way to elicit hyperparameters as in the paper. For both dead and live rats, all edges are considered not significant by the competitor. This



(a) (b) (c) (d)

Figure 2: Posterior mode correlation matrices (MAP estimation) for the edge detection problem from the dead rat (a,b) and live rat data (c,d). (a,c) first 7 regions for the competitor; (b) ours with $\rho_0 = 0, \nu = 29$; (d) ours with $\rho_0 = 0.3, \nu = 29$.

is because the variance of each edge is too large, possibly caused by covariance samples of poor quality-correlated, ill-conditioned, and based on small sample sizes. By contrast, our model is able to give reasonable results.

References

Achard, S. & Gannaz, I. (2016), 'Multivariate wavelet whittle estimation in long-range dependence', *Journal of Time Series Analysis* **37**(4), 476–512.

Acharyya, S., Zhang, Z., Bhattacharya, A. & Pati, D. (2023), 'Bayesian hierarchical modeling on covariance valued data', *Stat* **12**(1), e534.

Drton, M. & Perlman, M. D. (2007), 'Multiple testing and error control in gaussian graphical model selection', *Statist. Sci.* **22**(3).

Ginestet, C. E., Li, J., Balachandran, P., Rosenberg, S. & Kolaczyk, E. D. (2017), 'Hypothesis testing for network data in functional neuroimaging', *The Annals of Applied Statistics* pp. 725–750.

Gorbach, T., Lundquist, A., De Luna, X., Nyberg, L. & Salami, A. (2020), 'A Hierarchical Bayesian Mixture Model Approach for Analysis of Resting-State Functional Brain Connectivity: An Alternative to Thresholding', *Brain Connectivity* **10**(5), 202–211.

Jiang, Y., Chevaux, A., Arbel, J., Kon Kam King, G., Meiring, W., Petersen, A. & Achard, S. (2025), 'A bayesian approach for estimating brain functional connectivity networks in resting-state fmri data', *In preparation* .

Lazar, N. (2008), *The statistical analysis of functional MRI data*, Springer Science & Business Media.

Patel, R. S., Bowman, F. D. & Rilling, J. K. (2006), 'A Bayesian approach to determining connectivity of the human brain', *Human Brain Mapping* **27**(3), 267–276.

Vehtari, A., Gelman, A. & Gabry, J. (2017), 'Practical bayesian model evaluation using leave-one-out cross-validation and waic', *Statistics and Computing* **27**, 1413–1432.

Inférence de réseaux phylogénétiques par ABC-Forêts Aléatoires (ABC-RF) : application au processus de domestication du riz

Charles-Elie Rabier [1,2] & Jean-Christophe Glaszmann [3]

[1] *IMAG, Université de Montpellier, CNRS, France*
[2] *Univ Angers, Institut Agro, INRAE, IRHS, SFR QUASAV, F-49000 Angers, France, charles-elie.rabier@univ-angers.fr*
[3] *CIRAD, UMR AGAP, Montpellier, France, Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales (AGAP), Université de Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France, jean-christophe.glaszmann@cirad.fr*

Résumé. Nous nous intéressons dans cette étude, à l'inférence de réseaux phylogénétiques par ABC-Forêts Aléatoires (ABC-RF). Les réseaux phylogénétiques sont des graphes dirigés acycliques, qui peuvent représenter des scénarios évolutifs. Le champ d'application porte ici sur l'histoire de domestication du riz en Asie. Ce processus de domestication fait l'objet de débats permanents dans la communauté scientifique. Notre travail, basé sur des données de séquences génomiques, se décompose en trois étapes successives. On commence tout d'abord par explorer librement l'espace entier des réseaux phylogénétiques, par la méthode SnappNet dans un cadre Bayésien (MCMC), sur de multiples petits jeux de données. A noter que SnappNet présente l'avantage d'intégrer sur toutes les histoires locales, i.e. sur tous les arbres de locus. Suite aux nombreuses et diverses hybridations obtenues, nous constatons qu'il s'avère impossible de dégager un réseau phylogénétique faisant consensus entre jeux de données. Nous proposons dès lors, en deuxième étape, de nous concentrer sur 16 réseaux phylogénétiques plausibles, tout en augmentant le nombre de séquences. Avec l'aide de SnappNet dans un cadre de vraisemblance pénalisée, on extrait les 6 meilleurs réseaux. Enfin, afin de pouvoir voter pour un unique scénario en s'appuyant sur un jeu de données encore plus conséquent, nous présentons une étude ABC-RF basée sur des statistiques résumées inspirées par la littérature réseaux phylogénétiques. Nos statistiques résumées sont calculées sous le "*Network Multispecies Coalescent*" et implémentées au sein de notre simulateur SimSnappNet . Les distributions a priori sont calibrées à l'aide des estimations de paramètres obtenus à l'étape deux. Nous montrons qu'ABC-RF semble présenter un potentiel intéressant pour l'inférence de réseaux phylogénétiques sous le "*Network Multispecies Coalescent*".

Mots clés. ABC-Forêts Aléatoires, Réseaux phylogénétiques, Evolution, MCMC, Vraisemblance pénalisée, Statistique Bayésienne

Abstract. We focus here on phylogenetic network inference by means of the ABC-Random Forest method. The application field is the rice domestication process in Asia. We present here a methodological study based on a rich stochastic model, that incorporates introgression events, incomplete lineage sorting, and mutations that happen over time. Our study, relying on genomic data, consists in three steps. First, as a preliminary analysis, we explore freely the whole space of evolutionary scenarios, thanks to SnappNet in a Bayesian framework. Since it was impossible to draw a consensus phylogenetic network from the large diversity of inferred evolutionary histories, we propose in a second step, to concentrate on 16 rice evolutionary scenarios of interest. It enables us to handle more sequences simultaneously and to focus on a few domestication histories. With the help of SnappNet in a penalized

likelihood framework, we evaluate the likelihoods of the different phylogenetic networks and we select the best networks according to the AIC criteria. Last but not least, in a third step, we present an ABC-RF study, relying on an even larger data set and on summary statistics inspired by phylogenetic network literature. Our summary statistics are computed under the "*Network Multispecies Coalescent*", and implemented within our simulator SimSnappNet . Parameter values for the different prior distributions are chosen, thanks to SnappNet 's estimations at step two. This ABC-RF study allows us to discriminate among the six best scenarios found by SnappNet at step two.

Keywords. ABC-Random Forest, Phylogenetic networks, Evolution, Monte Carlo Markov Chain, Penalized likelihood, Bayesian statistics

1 Introduction et feuille de route

Les bouleversements impressionnants en biotechnologie ont généré un flot de données considérable permettant l'exploration systématique du vivant, à différentes échelles d'organisation et le recours croissant en biologie à la modélisation mathématique et aux sciences du numérique. Un défi actuel est de concevoir des méthodes statistiques, reposant sur des modèles biologiques complexes, tout en exploitant le volume considérable de données, comme par exemple les nombreuses séquences génomiques. Il s'avère désormais acquis que les données génomiques résultent d'une histoire évolutive riche, que se soit localement (marqueur SNP, ou gène) ou globalement (à l'échelle du génome). Au niveau local, on peut mentionner le tri de lignées incomplet (cf. Ebersberger et al., 2007) ou les transferts horizontaux de gènes (e.g. bactéries, fungi, Szollosi et al., 2015), deux évènements évolutifs, responsables de la discordance entre arbres de gènes et arbres d'espèces (Rokas et al, 2003). Au niveau de l'histoire globale des génomes, soulignons l'hybridation (assez fréquente chez les plantes), évènement évolutif qui intervient lorsque des individus appartenant à deux populations génétiquement différentes se reproduisent, créant une nouvelle population. Les approches basées sur les réseaux phylogénétiques (Huson et al., 2010) permettent la détection d'événements d'évolution réticulée, et donc la détection d'hybridation.

Bien que les méthodes proposées dans notre présente étude s'avèrent relativement génériques, on choisira comme champs applicatif, le processus de domestication du riz en Asie. Il existe principalement quatre types de riz cultivés, à savoir Japonica, Indica, circum Aus (cAus), et circum Basmati (cBasmati). Le processus de domestication du riz en Asie fait l'objet de perpétuels débats et plusieurs thèses s'opposent, à savoir une seule domestication (e.g. Huang et al., 2012, Wu et al., 2023), ou de multiples domestications (e.g. Civan et al., 2015, Jing et al., 2023). Afin d'approfondir la question de la domestication, notre étude dans sa globalité, se veut générale et sans a priori sur un scénario évolutif particulier.

Dans ce qui suit, on commencera par brièvement rappeler quelques méthodes statistiques de reconstruction de réseaux phylogénétiques. On s'attardera par la suite, sur le concept de calcul Bayésien approché (ABC) et d'ABC-Forêts Aléatoires (ABC-RF). Dans une deuxième partie, on décrira le modèle stochastique investi dans notre étude, et on présentera nos résultats.

2 Quelques méthodes statistiques de reconstruction de réseaux phylogénétiques

Les réseaux phylogénétiques (cf. Solis-Lemus et Ané, 2016) sont des graphes dirigés sans cycles dirigés et avec une seule racine, qui peuvent décrire les transferts horizontaux de gènes (e.g. bactéries), les phénomènes d'hybridations (e.g. plantes) ainsi que les introgressions (e.g. plantes et animaux). Les gènes sont transférés horizontalement lorsqu'ils sont acquis par une population à travers un processus autre que la reproduction. Les introgressions représentent l'intégration d'allèles d'une population dans une autre existante.

La reconstitution de réseaux phylogénétiques est un enjeu majeur car elle permet de répondre à de nombreuses questions soulevées par les biologistes. A l'heure actuelle, les méthodes de pseudo-vraisemblance et les méthodes Bayésiennes, toutes deux basées sur des modèles biologiques complexes, sont très populaires dans la reconstruction de réseaux phylogénétiques. Bien que les méthodes de pseudo vraisemblance soient approchées, elles présentent l'avantage de gérer de très gros jeux de données génomiques. Elles reposent sur deux étapes. Le réseau est tout d'abord divisé en sous réseaux avec un petit nombre de feuilles (i.e. espèces), puis la vraisemblance de chaque sous-réseau est calculée. Dans un deuxième temps, ces vraisemblances sont combinées afin d'obtenir une "pseudo-vraisemblance" du réseau phylogénétique entier (e.g. SNAQ de Solis-Lemus et Ané (2016) et Zhu et al. (2019)).

D'autre part, les méthodes Bayésiennes sont très prisées car elles fournissent une distribution de réseaux, pour lesquels on peut explorer l'incertitude sur les différents clade. La première méthode Bayésienne a été proposée par Wen et al. (2016). Par la suite, les méthodes SpeciesNetwork de Zhang et al. (2017) et MCMCBiMarkers de Zhu et al. (2018) ont été implémentées dans la plateforme beast 2 (Bouckaert et al., 2019) et dans PhyloNet (Wen et al., 2018), respectivement. Plus récemment, nous avons proposé la méthode SnappNet (Rabier et al., 2021) qui est basée exactement sur le même modèle mathématique que MCMCBiMarkers, excepté que les deux méthodes considèrent différentes distributions a priori pour le réseau. L'avantage de SnappNet sur MCMCBiMarkers, vient du fait que le calcul de vraisemblance est largement plus rapide avec SnappNet, de par le traitement mathématique. En particulier, SnappNet calcule des probabilités jointes, alors que MCMCBiMarkers est basée sur une approche combinatoire. La complexité du calcul de vraisemblance avec SnappNet est gouvernée par la scanwidth qui mesure le nombre maximal d'arêtes du réseau simultanément impliquées dans les calculs de probabilités jointes.

3 Le concept d'ABC-Forêts Aléatoires (ABC-RF)

Bien que SnappNet ait réduit considérablement les calculs de vraisemblance comparé à MCMCBiMarkers, le calcul de vraisemblance s'avére coûteux avec SnappNet dès que l'on traite de nombreux individus par espèces (cf. Rabier et al., 2021). Ainsi, afin de traiter plus de données, nous proposons ici une étude originale par ABC-RF, s'appuyant sur des statistiques résumées inspirées de la littérature des réseaux phylogénétiques. Le modèle stochastique investi demeure identique à celui de SnappNet.

Avant de s'attarder sur le concept d'ABC-RF, rappelons tout d'abord l'approche ABC. ABC, pour "Approximate Bayesian Computation", est une méthode computationnelle en statistique Bayésienne, qui permet d'approximer la distribution a posteriori. Contrairement aux méthodes Bayésiennes classiques comme SnappNet, les méthodes ABC (Rubin et al., 1984) présentent l'avantage de s'affranchir du calcul de vraisemblance (cf. Marin et al., 2012 pour une revue). Elles ont été motivées par la biologie (Tavaré et al., 1997, Pritchard et al., 1999) et sont très populaires en génétique des populations (e.g. Boitard et al., 2016, Jay et al., 2017). On recense également des applications en phylogénie (e.g. Saulnier et al., 2017, He et al., 2023).

Des approches ABC-Forêts Aléatoires (ABC-RF) ont été proposées par Pudlo et al. (2015), Raynal et al. (2019), Estoup et al. (2018) et ont déjà montré récemment tout leur potentiel applicatif (e.g. Chapuis et al., 2020, Gilabert et al., 2023). ABC-RF se distingue de l'ABC classique car elle ne prend pas en compte les distributions a posteriori traditionnellement utilisées. ABC-RF consiste à construire un classifieur sur la base du concept très célèbre des forêts aléatoires (Breiman, 2001) et d'une table de référence constituée d'un grand nombre de statistiques résumées. Par la suite, il est également possible d'approximer la probabilité a posteriori du modèle selectionné.

ABC-RF s'avère bien moins difficile qu'ABC à mettre en oeuvre en terme de calibration. ABC-RF nécessite beaucoup moins de simulations à partir de la loi a priori car elle ne repose pas sur une méthode d'acceptation/rejet : on peut ainsi remplir aisément une table de référence basée sur un très grand nombre de statistiques résumées.

4 Description du modèle stochastique investi

Le modèle stochastique investi dans notre étude est le modèle sous-jacent à SnappNet et à MCMCBiMarkers. C'est un modèle hiérarchique à plusieurs niveaux. Tout d'abord, le réseau phylogénétique reflète l'histoire réticulée gobale des espèces. Par la suite, l'arbre de gènes symbolise l'histoire locale à une position du génome (i.e. à un site). Il est modélisé via un processus de coalescence qui évolue à l'intérieur du réseau phylogénétique. Cela traduit ainsi le tri de lignées incomplet. Enfin, le long des branches de l'arbre de gènes, un processus Markovien représente les mutations de l'ADN intervenant au fil du temps. On propose ci-dessous une description détaillée des différents étapes.

4.1 Réseaux phylogénétiques

Un réseau phylogénétique est un graphe dirigé acyclique. Les branches du réseau sont dirigées de haut en bas, de manière à avancer dans le temps. En d'autres termes, les ancêtres les plus anciens sont situés en haut du réseau alors que les feuilles se situent en bas. Un réseau phylogénétique présente deux types de noeuds. Un noeud est soit a) binaire avec une arête entrante et deux arêtes sortantes, ou b) dit de réticulation avec deux arêtes entrantes et une arête sortante. Dans le modèle de SnappNet, chaque branche du réseau a sa propre longueur mesurée en nombre de mutation attendues, et chaque branche a son propre paramètre de taille

de population.

4.2 Modèle pour l'arbre de gènes

Pour chaque site, l'arbre de gènes est obtenu suivant le "*Network Multispecies Coalescent*" (Yu et al., 2012). Le processus débute aux feuilles du réseau phylogénétique et remonte dans le temps, jusqu'à ce que toutes les lignées coalescent. Au départ, la coalescence a lieu entre les lignées qui appartiennent à la même espèce. 2 lignées coalescent avec taux $2\mu/\theta$, où μ and θ désignent respectivement un taux de mutation et un paramètre de taille de population propre à l'espèce. Si l'on assume que cette espèce est composée de k lignées, le premier temps de coalescence suit une loi exponentielle $\mathcal{E}\left(k(k-1)\mu/\theta\right)$, car la coalescence de chaque combinaison de 2 lignées est équiprobable. Lorsque $k = 2$, l'espérance du temps de coalescence est $\theta/(2\mu)$. θ est le nombre moyen de mutations, séparant 2 individus. Lorsque les lignées (qui n'ont pas pu coalescer) remonte à un noeud de spéciation délimitant deux espèces, les lignées sont regroupées et ont possibilité de coalescer le long de la nouvelle branche, en tenant compte de la valeur de θ associée à cette branche. Lorsque les lignées (qui n'ont pas pu coalescer) remontent à un noeud de réticulation, chaque lignée choisit indépendamment son origine ancestrale (modèle de Yu et al., 2011), selon une loi de Bernoulli de paramètre γ.

Pour finir, au dessus de la racine du réseau phylogénétique, la coalescence intervient entre toutes les lignées et et le processus se termine lorsqu'uniquement une seule lignée ancestrale est restante. Comme son cousin Snapp (Bryant et al., 2012), SnappNet considère $\mu = 1$ (cf. explications ci-dessous).

4.3 Modèle de mutation

Comme Snapp, SnappNet considère des marqueurs bialléliques et les couleurs *rouges* et *vertes* représentent les deux allèles. Les marqueurs évoluent le long des branches des arbres de gènes, selon une chaine de Markov à temps continu, où u et v dénotent respectivement les taux instantanées de muter de rouge à vert, et de vert à rouge. Sous ce modèle, pour une branche de longueur T, il y a en moyenne $2uvT/(u+v)$ mutations. Alors, le taux de mutation μ est de $2uv/(u+v)$. Imposer la contrainte $2uv/(u+v) = 1$ (i.e. $\mu = 1$), permet de mesurer les longueurs de branche en taux de substitutions par site (i.e. distance génétique).

5 Nos résultats

Notre étude est composée de trois étapes.

On commence tout d'abord par explorer librement l'espace entier des scénarios évolutifs, par la méthode SnappNet dans un cadre Bayésien, sur de multiples petits jeux de données. Pour l'analyse par MCMC, sur chaque jeu de données, deux chaines de Markov ont été considérées pour chaque échantillonnage de 12,000 marqueurs le long du génome. Suite aux nombreuses et diverses hybridations obtenues, nous constatons qu'il s'avère impossible de dégager un réseau phylogénétique faisant consensus entre jeux de données.

Nous proposons dès lors, en deuxième étape, de nous concentrer sur 16 scénarios évolutifs. Ainsi, en nous focalisant uniquement sur quelques scénarios, nous sommes en mesure de gérer plus de séquences simultanément. La plupart de ces 16 scénarios demeurent plausibles pour l'histoire évolutive du riz en Asie. Avec l'aide de SnappNet dans un cadre de vraisemblance pénalisée, on se propose d'évaluer les vraisemblances des 16 réseaux phylogénétiques puis de sélectionner les meilleurs réseaux selon le critère AIC. On illustre par la même occasion le concept de scanwidth qui gouverne la complexité de SnappNet. En effet, les réseaux considérés présentent un, deux, ou trois noeuds de réticulation. A noter que SnappNet présente l'avantage d'intégrer sur tous les arbres gènes dans son calcul de vraisemblance.

Pour finir, afin d'être en mesure d'explorer un plus grand jeu de données, nous présentons une étude ABC-RF basée sur des statistiques résumées inspirées par la littérature réseaux phylogénétiques. Nos statistiques résumées sont calculées sous le "*Network Multispecies Coalescent*" (Yu et al., 2012) et implémentées au sein de notre simulateur SimSnappNet (Rabier et al., 2021). Les valeurs des paramètres associés aux différentes distribution a priori pour ABC-RF sont choisies grâce aux estimation fournies par SnappNet à l'étape deux. Au final, nous avons considéré 8 combinaisons de distributions a priori, un total de 562 statistiques résumées et 84,000 simulations pour chaque combinaison. Dans tous les cas testés, le taux d'erreur a priori s'est avéré relativement faible pour chacun des scénarios, illustrant les bonnes performances du classifieur. Ainsi, l'ABC-RF couplé à un choix judicieux de statistiques résumées semble présenter un fort potentiel dans le cadre de réseaux phylogénétiques et du tri de lignées incomplet. Par la suite, sur les données réelles de riz, on montre qu'ABC-RF nous permet d'extraire le meilleur des six premiers scénarios suggérés par SnappNet à l'étape deux. Le scénario fortement retenu opte pour une seule domestication, celle de Japonica. Japonica aurait contribué par la suite à la domestication d'Indica et de cAus.

Bibliographie

Civan, P., Craig, H., Cox, C. J., Brown, T. A. (2015). Three geographically separate domestications of Asian rice. *Nature plants*, 1, 15164.

Marin, J. M., Pudlo, P., Robert, C. P., Ryder, R. J. (2012), Approximate Bayesian computational methods *Statistics and Computing*, 22(6), 1167-1180.

Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J. M., Gautier, M., Robert, C. P. (2015). Reliable ABC model choice via random forests. *Bioinformatics*, 32(6), 859-866.

Rabier CE, Berry V, Stoltz M, Santos JD, Wang W, Glaszmann JC, Pardi F, Scornavacca C (2021). On the inference of complex phylogenetic networks by Markov Chain Monte-Carlo. *PLoS Computational Biology*. 17(9), e1008380.

Solís-Lemus, C., Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting, *PLoS genetics*, 12(3), e1005896.

Statistique bayésienne computationnelle

Inférence variationnelle pour l'approximation de priors de référence à l'aide de réseaux de neurones

Nils Baillie[1,2] & Antoine Van Biesbroeck[1,2] & Clément Gauchy[3]

[1] *Université Paris-Saclay, CEA, Service d'Études Mécaniques et Thermiques, 91191 Gif-sur-Yvette, France*
nils.baillie@cea.fr
[2] *CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France*
nils.baillie@polytechnique.edu
[3] *Université Paris-Saclay, CEA, Service de Génie Logiciel pour la Simulation, 91191 Gif-sur-Yvette, France*

Résumé. En statistique bayésienne, le choix du prior peut avoir une influence importante sur la loi a posteriori et l'estimation des paramètres. Afin de limiter la subjectivité issue de l'information a priori, on peut utiliser le cadre des priors de référence. Cependant, calculer de tels priors est une tâche difficile en général. Nous présentons dans la suite un algorithme flexible basé sur l'inférence variationnelle qui calcule une approximation de prior de référence à partir d'un ensemble de lois paramétriques fixé par un réseau de neurones. Nous montrons également que notre algorithme est capable de retrouver des priors de référence quand des contraintes sont spécifiées afin d'assurer la propreté de la solution. Nous proposons une approche simple pour retrouver une approximation pertinente de la loi a posteriori en utilisant des méthodes de Monte Carlo par chaînes de Markov (MCMC), y compris lorsque la densité du prior paramétrique est inconnue. La performance de l'algorithme est évaluée sur les distributions a priori et a posteriori.

Mots-clés. Priors de référence, Inférence variationnelle, Réseaux de neurones

Abstract. In Bayesian statistics, the choice of the prior can have a significant influence on the posterior and the parameter estimation. To limit the subjectivity introduced by prior information, one can use the framework of reference priors. However, computing such priors is a difficult task in general. We propose a flexible algorithm based on variational inference which computes approximations of reference priors from a set of parametric distributions using neural networks. We also show that our algorithm can recover reference priors under constraints that ensure the solution remains proper. We propose a simple approach to recover a relevant approximation of the posterior distribution using Markov Chain Monte Carlo (MCMC) methods, even when the prior density is unknown. The performance of the algorithm is evaluated on both prior and posterior distributions.

Keywords. Reference priors, Variational inference, Neural networks

1 Introduction

L'approche bayésienne en inférence statistique a pour but de produire des estimations à partir de la distribution a posteriori. Cette dernière est obtenue en mettant à jour la loi a priori avec les donnés observées grâce au théorème de Bayes. Cependant, la forme du posterior peut être lourdement influencée par le choix du prior lorsque la quantité de données disponibles est limitée ou quand le prior est très informatif. Pour cette raison, le choix du prior reste une tâche ardue qui doit être menée avec soin.

Dans ce travail, nous allons nous focaliser sur la théorie des priors de référence, introduite par Bernardo (1979) et formalisée par la suite par Berger et al. (2009). La principale motivation derrière la théorie des priors de référence est la maximisation de l'information apportée par les données lors de l'inférence bayésienne. Plus spécifiquement, les priors de référence (RP) sont construits pour maximiser la métrique d'information mutuelle, qui est définie comme une divergence entre le RP et le posterior qui en découle. Ainsi, cela permet aux données de jouer un rôle prédominant dans le cadre bayésien. Les RP sont reconnus pour leur nature objective dans des études pratiques (Van Biesbroeck (2024a)), cependant, ils ont le défaut d'être coûteux à calculer, d'autant plus que la dimension du problème augmente.

Différentes approches existent pour approximer ou échantillonner des distributions de probabilité dans le cadre bayésien, nous adoptons celle de l'inférence variationnelle. L'objectif est d'approcher une certaine distribution p en optimisant sur un ensemble de distributions paramétriques q_λ, c'est-à-dire de trouver la distribution q_{λ^*} qui est la meilleure approximation de p au sens de la minimisation d'une divergence, comme celle de Kullback-Leibler (KL) par exemple. Les travaux de Nalisnick & Smyth (2017) et Gauchy et al. (2023) ont entamé l'application de cette approche aux RP, notre apport est la généralisation de ce cadre à une large gamme de divergences, notamment les α-divergences, mais aussi à l'ajout de contraintes dans le problème d'optimisation. Ces dernières pouvant assurer la propreté du prior solution selon les résultats de Van Biesbroeck (2024b), ce qui vient répondre à un écueil courant des RP, à savoir leur impropreté. Les détails vis-à-vis de la construction de l'algorithme et l'ensemble des applications numériques sont disponibles dans la version complète de l'article (Baillie et al. (2025)).

2 Théorie des priors de référence

Nous nous intéressons à la situation où nous observons des échantillons de données i.i.d. $\mathbf{X} = (X_1, ..., X_N) \in \mathcal{X}^N$ où $\mathcal{X} \subset \mathbb{R}^d$. On suppose que la fonction de vraisemblance est connue : $L_N(\mathbf{X} \,|\, \theta) = \prod_{i=1}^N L(X_i \,|\, \theta)$ et $\theta \in \Theta \subset \mathbb{R}^q$ est le paramètre à inférer. On note π le prior sur θ et $p_{\pi,N}(\mathbf{X}) = \int_\Theta \pi(\theta) L_N(\mathbf{X} \,|\, \theta) d\theta$ est la loi marginale. On considère une définition généralisée de l'information mutuelle proposée par Van Biesbroeck (2024a) :

$$I_{\mathrm{D}_f}(\pi; L_N) = \int_\Theta \mathrm{D}_f(p_{\pi,N} || L_N(\cdot \,|\, \theta)) \pi(\theta) d\theta,$$

avec la f-divergence $\mathrm{D}_f(p \,||\, q) = \int_\Omega f\,(p/q)\, q d\omega$ (f étant une fonction convexe telle que $f(1) = 0$). L'information mutuelle classique considérée par Berger et al. (2009) correspond au cas

où $f = -\log$. Il a été démontré par Clarke & Barron (1994) et Van Biesbroeck (2024a) que le prior qui maximise l'information mutuelle dans le cas limite $N \to +\infty$ est le prior de Jeffreys, respectivement pour l'information mutuelle classique ($f = -\log$) et l'information mutuelle généralisée avec les α-divergences ($f = f_\alpha$) où:

$$f_\alpha(x) = \frac{x^\alpha - \alpha x - (1-\alpha)}{\alpha(\alpha-1)}, \quad \alpha \in (0,1).$$

Le prior de Jeffreys est défini comme étant $J(\theta) \propto \det(\mathcal{I}(\theta))^{1/2}$ où $\mathcal{I}(\theta)$ est la matrice d'information de Fisher. Afin d'évaluer les performances de notre algorithme, il nous faut pouvoir comparer le prior en sortie de notre code de calcul et le prior théorique, seulement, le prior de Jeffreys a tendance a être impropre dans de nombreuses situations. Cela rend une comparaison rigoureuse impossible, c'est pourquoi nous adaptons l'algorithme pour qu'il puisse résoudre une version contrainte du problème d'optimisation initial :

$$\tilde{\pi}^* \in \arg\max_{\pi \text{ prior}} I_{\mathrm{D}_{f_\alpha}}(\pi; L_N) \quad \text{avec} \quad \mathcal{C}(\pi) < \infty,$$

où $\mathcal{C}(\pi)$ est une contrainte de la forme $\int_\Theta a(\theta)\pi(\theta)d\theta$, a étant une fonction positive. Lorsqu'une α-divergence est considérée dans l'information mutuelle et que a vérifie :

$$\int_\Theta J(\theta)a(\theta)^{1/\alpha}d\theta < \infty \quad \text{et} \quad \int_\Theta J(\theta)a(\theta)^{1+(1/\alpha)}d\theta < \infty,$$

alors Van Biesbroeck (2024b) a démontré que le RP contraint $\tilde{\pi}^*$ prend asymptotiquement la forme $\tilde{\pi}^*(\theta) \propto J(\theta)a(\theta)^{1/\alpha}$, qui est un prior propre.

3 Approximation variationnelle du prior de référence

Nous considérons un ensemble paramétrique de priors : $\{\pi_\lambda, \lambda \in \Lambda\}$, $\Lambda \subset \mathbb{R}^L$. On cherche alors à trouver $\arg\max_{\lambda\in\Lambda} I_{\mathrm{D}_f}(\pi_\lambda; L_N)$. Nous reprenons l'approche de Gauchy et al. (2023) qui consiste à définir π_λ implicitement:

$$\theta \sim \pi_\lambda \iff \theta = g(\lambda, \varepsilon) \quad \text{où} \quad \varepsilon \sim \mathbb{P}_\varepsilon.$$

Ici, g représente un réseau de neurones paramétré par λ, on suppose donc que g est dérivable en λ. La variable ε peut être vue comme une variable latente, elle admet une distribution $\mathbb{P}_\varepsilon$ simple à échantillonner avec une densité connue, en pratique, on utilise une loi normale multivariée $\mathcal{N}(0, \mathbb{I}_{p\times p})$. Cette construction nous permet de considérer une vaste famille de priors, cependant, il faut mentionner que la densité de π_λ n'est pas connue, sauf dans des cas très simples. On a seulement accès à des échantillons $\theta \sim \pi_\lambda$.

Cette approche implicite semble être la plus prometteuse par rapport à des méthodes de calcul direct (Berger et al. (2009)) ou des méthodes basées sur les chaînes de Markov selon les comparaisons effectuées par Nalisnick & Smyth (2017) sur des cas uni-dimensionnels.

Afin de résoudre numériquement ce problème d'optimisation, notre algorithme se base sur l'optimisation par gradient stochastique avec une fonction objectif correspondant à une borne

inférieure de l'information mutuelle. Cette fonction a été introduite par Nalisnick & Smyth (2017) mais nous proposons une généralisation aux f-divergences avec f décroissante:

$$B_{\mathrm{D}_f}(\pi; L_N) = \int_\Theta \int_{\mathcal{X}^N} f\left(\frac{L_N(\mathbf{X} \mid \hat{\theta}_{MLE})}{L_N(\mathbf{X} \mid \theta)}\right) \pi(\theta) L_N(\mathbf{X} \mid \theta) d\mathbf{X} d\theta,$$

dont le gradient peut s'écrire ainsi :

$$\frac{\partial B_{\mathrm{D}_f}}{\partial \lambda_l}(\pi_\lambda; L_N) = \mathbb{E}_\varepsilon \left[\sum_{j=1}^{q} \frac{\partial \tilde{B}}{\partial \theta_j}(g(\lambda, \varepsilon)) \frac{\partial g_j}{\partial \lambda_l}(\lambda, \varepsilon)\right],$$

où:

$$\frac{\partial \tilde{B}}{\partial \theta_j}(\theta) = \mathbb{E}_{\mathbf{X} \sim L_N(\cdot \mid \theta)} \left[\frac{\partial \log L_N}{\partial \theta_j}(\mathbf{X} \mid \theta) F\left(\frac{L_N(\mathbf{X} \mid \hat{\theta}_{MLE})}{L_N(\mathbf{X} \mid \theta)}\right)\right], \quad \text{avec} \quad F(x) = f(x) - x f'(x).$$

En pratique, le gradient est approximé en utilisant des estimateurs de Monte Carlo basés sur des échantillons de $\varepsilon \sim \mathbb{P}_\varepsilon$ et de $\mathbf{X} \sim \mathbb{P}_{\mathbf{X}|\theta}$. Il est important de séparer le calcul du gradient de g, calculé par différentiation automatique (PyTorch), et celui de $\tilde{B}$, obtenu par échantillonnage, car ce dernier peut rendre la différentiation automatique instable. L'optimisation est ensuite faite via l'emploi de l'optimiseur Adam.

Dans le cas de figure où une contrainte est imposée dans le problème d'optimisation :

$$\arg\max_{\lambda \in \Lambda} \mathcal{B}_{\mathrm{D}_f}(\pi_\lambda; L_N) \quad \text{avec} \quad \mathcal{C}(\pi_\lambda) := \mathbb{E}_{\theta \sim \pi_\lambda}[a(\theta)] - b = 0,$$

on utilise le même algorithme, mais avec le lagrangien augmenté :

$$\mathcal{L}_A(\lambda, \eta, \tilde{\eta}) = \mathcal{B}_{\mathrm{D}_f}(\pi_\lambda; L_N) + \eta \mathcal{C}(\pi_\lambda) - \frac{\tilde{\eta}}{2} \mathcal{C}(\pi_\lambda)^2.$$

La valeur de b à utiliser est obtenue en approximant les deux intégrales à la fin de la partie 2, ce qui nécessite un calcul préliminaire du prior de Jeffreys.

Afin de construire des estimateurs a posteriori, nous proposons une méthode d'échantillonnage de θ selon sa loi a posteriori $\pi_\lambda^{\mathrm{post}}(\theta \mid \mathbf{X})$. Cependant la distribution a priori π_λ est inconnue car seulement implicite en général. Nous construisons alors nos échantillons à partir de la loi a posteriori de la variable ε dont la densité est connue à une constante près:

$$\pi_{\varepsilon,\lambda}^{\mathrm{post}}(\varepsilon \mid \mathbf{X}) \propto p_\varepsilon(\varepsilon) L_N(\mathbf{X} \mid g(\lambda, \varepsilon)) \; ; \quad \text{si } \varepsilon \sim \pi_{\varepsilon,\lambda}^{\mathrm{post}}(\varepsilon \mid \mathbf{X}) \text{ alors } g(\lambda, \varepsilon) \sim \pi_\lambda^{\mathrm{post}}(\theta \mid \mathbf{X}).$$

On utilise alors une variante adaptative de Metropolis-Hastings pour réaliser cet échantillonnage.

4 Application numérique

Nous présentons ici un exemple d'utilisation de l'algorithme sur un modèle jouet où on veut inférer la variance d'une loi normale. On a $\mathbf{X} = (X_1, ..., X_N) \in \mathbb{R}^N$ où $X_i \sim \mathcal{N}(0, \theta)$. On

utilise $p = 10$, $N = 10$, $U = 1000$ et $T = 50$. Cependant, lorsqu'aucune contrainte n'est imposée, le prior de Jeffreys $J(\theta) \propto 1/\theta$ est impropre, mais le posterior qui en découle est une loi inverse-gamma propre.

Afin de pouvoir réaliser une comparaison sur les priors, on ajoute une contrainte définie par la fonction $a(\theta) = 1/(\theta^{-1} + \theta)$, avec $\alpha = 0.5$, on a $b = \pi/8$. Le RP sous contrainte est propre et a une densité explicite, il en est de même pour le posterior contraint.

Nous optimisons dans cet exemple un réseau de neurones à une couche linéaire avec une activation softplus dans le cas non contraint et exponentielle dans le cas contraint. Les résultats suivants sont obtenus avec la fonction borne inférieure $\mathcal{B}_{\mathrm{D}_{f_\alpha}}$ avec α-divergence en prenant $\alpha = 0.5$. Pour la distribution a posteriori, le jeu de données consiste en 10 échantillons de loi normale avec $\theta_{true} = 1$.

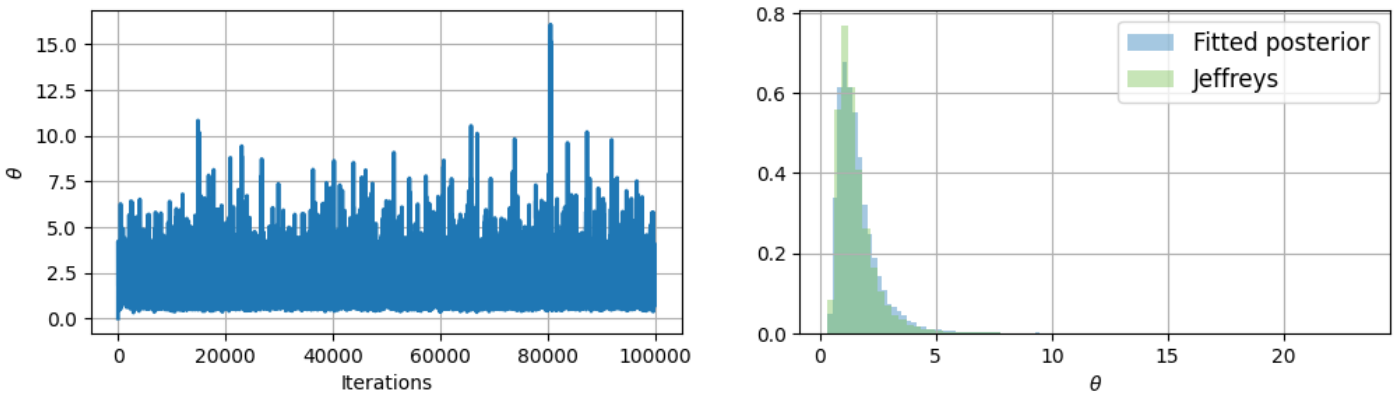
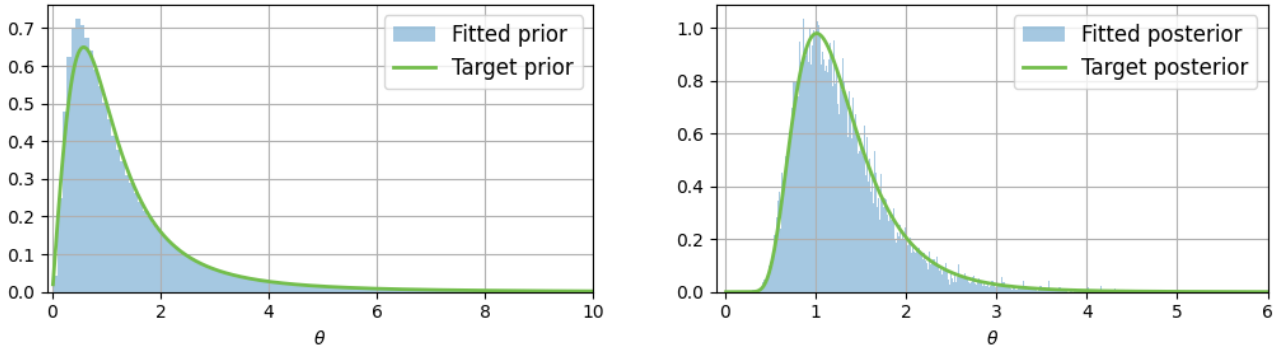


Figure 1: Gauche : Chaîne de Markov au cours des itérations, le temps de chauffe utilisé est $5 \cdot 10^4$. Droite : Histogrammes du posterior paramétrique sans contrainte et du posterior de Jeffreys (inverse-gamma), tous deux obtenus avec $5 \cdot 10^4$ échantillons.



(a) Priors

(b) Posteriors

Figure 2: Comparaison de l'histogramme du prior (resp. posterior) paramétrique sous contrainte obtenu à partir de 10^6 (resp. $5 \cdot 10^4$) échantillons avec la densité de la loi cible.

On peut observer grâce aux figures précédentes que notre algorithme est capable de retrou-

ver une approximation pertinente de la loi a posteriori dans le cas sans contrainte, et des deux lois, prior et posterior, dans le cas avec contrainte.

5 Conclusion

Nous avons développé un algorithme d'inférence variationnelle permettant de retrouver le prior de référence pour une forme généralisée de l'information mutuelle avec f-divergences. De plus, nous avons adapté cette approche au cas où des contraintes seraient imposées sur le prior pour assurer la propreté du posterior. Notre implémentation est open-source et flexible, et peut être utilisée pour une large gamme de modèles statistiques.

Parmi les perspectives, on peut mentionner l'amélioration de l'approximation des queues de distribution, qui est un problème complexe en inférence variationnelle, ainsi que la stabilité de l'algorithme lorsque des réseaux profonds sont utilisés.

Remerciements : Ce travail de recherche a été financé par le CEA et l'institut SEISM (https://www.institut-seism.fr/en/).

References

Baillie, N., Van Biesbroeck, A., & Gauchy, C. (2025). Variational inference for approximate reference priors using neural networks. https://arxiv.org/abs/2502.02364.

Berger, J. O., Bernardo, J. M., & Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, 37(2), 905 – 938. doi:10.1214/07-AOS587.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B*, 41(2), 113–147. doi:10.1111/j.2517-6161.1979.tb01066.x.

Clarke, B. S. & Barron, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1), 37–60. doi:10.1016/0378-3758(94)90153-8.

Gauchy, C., Van Biesbroeck, A., Feau, C., & Garnier, J. (2023). Inférence variationnelle de lois a priori de référence. Dans *Proceedings des 54èmes Journées de Statistiques (JdS)*. https://jds2023.sciencesconf.org/resource/page/id/19.

Nalisnick, E. & Smyth, P. (2017). Learning approximately objective priors. Dans *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*. doi:10.48550/arXiv.1704.01168.

Van Biesbroeck, A. (2024a). Generalized mutual information and their reference priors under csizar f-divergence. arXiv. https://arxiv.org/abs/2310.10530.

Van Biesbroeck, A. (2024b). Properly constrained reference priors decay rates for efficient and robust posterior inference. arXiv. https://arxiv.org/abs/2409.13041.

Approximate Bayesian Computation with Deep Learning and Conformal prediction

Meili Baragatti [1] & Céline Casenave [2] & Bertrand Cloez [2] & David Métivier [2] & Isabelle Sanchez [2]

[1] *MISTEA, Université de Montpellier, INRAE, Institut Agro, Montpellier, France*
[2] *MISTEA, Université de Montpellier, INRAE, Montpellier, France*

Résumé. Les méthodes de calcul bayésien approximatif (ABC) sont couramment utilisées pour approximer les distributions a posteriori dans les modèles dont les vraisemblances sont inconnues ou difficiles à calculer. Les méthodes ABC classiques sont basées sur des algorithmes de type "plus proche voisin" et reposent sur le choix de statistiques résumées, de distances entre jeux de données et d'un seuil de tolérance. Récemment, des méthodes combinant l'ABC avec des algorithmes d'apprentissage automatique plus complexes ont été proposées pour atténuer l'impact des choix de l'utilisateur. Dans cette présentation je présenterai la première méthode ABC, à notre connaissance, totalement exempte de statistiques résumées, de distance et de seuil de tolérance. En outre, contrairement aux méthodes ABC classiques, elle associe un intervalle de confiance (ayant une couverture marginale fréquentiste appropriée) à l'estimation de la moyenne a posteriori (ou à d'autres estimations de type moment).

Notre méthode, appelée ABCD-Conformal, utilise un réseau de neurones avec Monte Carlo Dropout pour fournir une estimation de la moyenne a posteriori (ou d'autres fonctions de type moment), et la théorie conformelle pour obtenir des ensembles de confiance associés. Je comparerai ses performances avec d'autres méthodes ABC sur plusieurs exemples, et vous montrerai qu'elle est efficace pour estimer des paramètres multidimensionnels, tout en étant "amortissable".

Mots-clés. Inférence sans vraisemblance, Approximate Bayesian computation, réseaux de neurones convolutionnels, dropout, prédiction conformelle, probability matching criterion.

Abstract. Approximate Bayesian Computation (ABC) methods are commonly used to approximate posterior distributions in models with unknown or computationally intractable likelihoods. Classical ABC methods are based on nearest neighbor type algorithms and rely on the choice of so-called summary statistics, distances between datasets and a tolerance threshold. Recently, methods combining ABC with more complex machine learning algorithms have been proposed to mitigate the impact of these "user-choices". In this talk, we propose what is, to our knowledge, the first ABC method that is completely free of summary statistics, distance, and tolerance threshold. Moreover, in contrast with usual generalizations of the ABC method, it associates a confidence interval (having a proper frequentist marginal coverage) with the posterior mean estimation (or other moment-type estimates).

Our method, named ABCD-Conformal, uses a neural network with Monte Carlo Dropout to provide an estimation of the posterior mean (or other moment type functionals), and conformal theory to obtain associated confidence sets. Being efficient for estimating multidimensional parameters and amortized, we test this new method on four different applications and compare it with other ABC methods from the literature.

Keywords. Likelihood-free inference, Approximate Bayesian computation, Convolutional neural networks, Dropout, Conformal prediction, probability matching criterion.

1 Introduction

In many situations, the likelihood function plays a key role in both frequentist or Bayesian inference. However, it is not unusual that such likelihood is difficult to work with either because no closed form expression exists, e.g., the corresponding statistical model is too complex, or because its computational cost is prohibitive. For example, the first situation occurs with max-stable processes [de Haan, 1984] or Gibbs random fields [Besag, 1974; Grelaud et al., 2009], while the second case frequently occurs with mixture models [Simola et al., 2021].

In this talk, we will restrict our attention to parameter inference in the Bayesian framework and in particular to Approximate Bayesian Computation (ABC), which is one of the most popular likelihood-free parameter inference method. Indeed, ABC can be an efficient alternative to overcome the situation of intractable likelihood functions [Marin et al., 2012]. Within the ABC framework, the need for computing the likelihood function is substituted by the need to be able to simulate from the statistical model under consideration. The first proposed ABC methods enabled to obtain an approximation of the posterior distribution of interest, from simulations [Pritchard et al., 1999]. The simplest sampler is simply an accept-reject algorithm, and will be denoted by "standard ABC" in the following. This vanilla sampler can be quite inefficient, as numerous draws can be rejected before accepting one. Alternative samplers have therefore been proposed in a search for computational efficiency, using MCMC or filtering algorithms (see Marjoram et al. [2003], Sisson et al. [2007], Baragatti and Pudlo [2014] and references therein). In particular, the Sequential Monte Carlo ABC (ABC-SMC, Sisson et al. [2007]; Beaumont et al. [2009]; Del Moral et al. [2012]) has become the gold standard in ABC methods. Even if more efficient than the standard ABC, these early ABC samplers are generally difficult to use when simulating complex systems or inferring high-dimensional parameters. This is why, with the increasing complexity of the data and models used nowadays, ABC samplers are still an active field of research. For instance, in order to improve the ABC-SMC sampler, guided approaches have been proposed, with the goal to rapidly target the region of the parameter space that is compatible with the observed data [Picchini and Tamborrino, 2024]. However, for most of these ABC methods, the degree of approximation of the posterior distribution depends strongly on the choices of the tolerance threshold, the distance and of the so-called summary statistics used during the ABC estimation. Unfortunately, identifying the most relevant summary statistics is problem dependent and in many situations the statistics are chosen by the practitioner, even if several studies provide guidance in this choice, see for instance Joyce and Marjoram [2008], McKinley et al. [2009], Blum et al. [2012], Fearnhead and Prangle [2012] among others.

Recently, several approaches managed to get rid of the tolerance level or to mitigate the need of identifying relevant summary statistics. Most of these approaches combine the ABC framework with machine learning methods, such as random forests or neural networks. For

instance, Raynal et al. [2018] propose an automatic selection of relevant summary statistics. They propose to consider a large set of potentially relevant summary statistics and use the random forest framework as a black box to get some point estimates of the posterior distribution given these summary statistics. They do not need any tolerance level, like the previously cited Simulation Based Inference (SBI) methods [Papamakarios and Murray, 2016; Papamakarios et al., 2019], which are called "ϵ-free" by their authors and use neural networks. Another approach of interest is the one of Akesson et al. [2022], which uses the estimated posterior mean as a summary statistic, this posterior mean being estimated through a Convolutional Neural Network (CNN). Consequently, since the emergence of ABC methods, a wide variety of methods have been proposed, from the standard "basic" ABC to methods using random forests or neural networks.

To quantify the uncertainties, Bayesian methods provide posterior credible intervals from the posterior distribution, through posterior quantiles or highest posterior density regions. This can be justified because, in many simple parametric models, the Bernstein-von Mises theorem ensures that posterior credible intervals are asymptotically as well, reconciling the frequentist and Bayesian approaches. However, these credible intervals are not necessarily relevant in practice. Indeed, note for instance that in nonparametric and high-dimensional models, research is still ongoing to understand better the link between the coverage of credible intervals and the coverage of frequentist, see Rousseau and Szabo [2016]; Datta et al. [2000]; Hoff [2023] among others. Indeed, the posterior coverage probability of a Bayesian credible interval does not always match the corresponding frequentist coverage probability. Quite often, they can even diverge in nontrivial ways, see for instance Wasserman [2011]. In this case, we say that the *probability matching criterion* is not verified.

This criterion can be hard to check when the exact posterior distribution is known, and even more so when the posterior is approximated [Frazier et al., 2018], which is the cases of most applications requiring an ABC method. Since the posterior distribution is inherently approximated by ABC methods, it is then crucial to account for the approximation error introduced by such algorithms.

2 Our contribution

In this talk, a new ABC implementation that does not require summary statistics, tolerance threshold and distance will be presented. Our method produces point estimates of the posterior distribution e.g., mean, variance, quantiles, without aiming to approximate the whole posterior distribution. It also provides associated confidence sets of the estimates. The estimates are obtained using neural networks with Monte Carlo Dropout, while the confidence sets are obtained through conformal prediction. Our ABC estimation method only requires a 1) sampling function, 2) a neural network (NN) whose architecture is tailored for the problem of interest e.g., CNN for images, GNN for data with a graph structure, RNN for sequential data. Furthermore, to obtain confidence intervals we need 3) a confidence level δ. While 1) is common to all ABC methods, requirement 2) is more novel yet still easy to implement, as well-known architectures can be used with minimal changes, i.e., the inclusion of Dropout (or concrete Dropout) layers. Another originality of this work is that it produces directly

confidence sets with proper marginal coverage, while usual ABC methods enable to obtain credible sets using δ-quantiles of the estimated posterior distribution.

3 Description of the talk

The talk will be organized as follows:

First the motivation for this work will be explained.

Then, the ABCD-Conformal methodology will be presented, combining neural networks and conformal prediction in a likelihood-free framework.

Eventually, the method will be illustrated and compared to other ABC methods, on several examples, before a discussion.

References

Akesson, M., Singh, P., Wrede, F., and Hellander, A. (2022). Convolutional neural networks as summary statistics for approximate bayesian computation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(06):1–1.

Baragatti, M. and Pudlo, P. (2014). An overview on approximate bayesian computation*. *ESAIM: Proc.*, 44:291–299.

Beaumont, M., Cornuet, J., Marin, J. M., and Robert, C. (2009). Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990.

Besag, J. (1974). Spatial interactions and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B. Statistical methodology*, 148:1–36.

Blum, M., Nunes, M., Prangle, D., and Sisson, S. (2012). A comparative review of dimension reduction methods in approximate bayesian computation. stat sci 28: 189-208. *Statistical Science*, 28.

Datta, G. S., Ghosh, M., Mukerjee, R., and Sweeting, T. J. (2000). Bayesian prediction with approximate frequentist validity. *The Annals of Statistics*, 28(5):1414 – 1426.

de Haan, L. (1984). A spectral representation for max-stable processes. *The Annals of Probability*, 12(4):1194–1204.

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, 22:1009–1020.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.

Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika*, 105(3):593–607.

Grelaud, A., Marin, J.-M., Robert, C. P., Rodolphe, F., and Taly, J.-F. (2009). ABC likelihood-free methods for model choice in gibbs random fields. *Bayesian Analysis*, 4.

Hoff, P. (2023). Bayes-optimal prediction with frequentist coverage control. *Bernoulli*, 29(2):901 – 928.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and bayesian computation. *Statistical applications in genetics and molecular biology*, 7:Article26.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate bayesian computational methods. *Statistics and computing*, 22(6):1167–1180.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the USA*, 100(26):15324–15328.

McKinley, T., Cook, A. R., and Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5(1).

Papamakarios, G. and Murray, I. (2016). Fast ϵ-free inference of simulation models with bayesian conditional density estimation.

Papamakarios, G., Sterratt, D., and Murray, I. (2019). Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 837–848. PMLR.

Picchini, U. and Tamborrino, M. (2024). Guided Sequential ABC Schemes for Intractable Bayesian Models. *Bayesian Analysis*, pages 1 – 32.

Pritchard, J., Seielstad, M., Perez-Lezaun, A., and Feldman, M. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16:1791–1798.

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., and Estoup, A. (2018). ABC random forests for Bayesian parameter inference. *Bioinformatics*, 35(10):1720–1728.

Rousseau, J. and Szabo, B. (2016). Asymptotic frequentist coverage properties of bayesian credible sets for sieve priors in general settings. *Annals of Statistics*, 48.

Simola, U., Cisewski-Kehe, J., and Wolpert, R. L. (2021). Approximate bayesian computation for finite mixture models. *Journal of Statistical Computation and Simulation*, 91(6):1155–1174.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765.

Wasserman, L. (2011). Frasian inference. *Statistical Science*, 26(3):322–325.

Améliorer la Simulation-Based Inference grâce à l'Approximate Bayesian Computation

Antoine Luciano [1] & Ryan Giordano [2]

[1] *CEREMADE, Université Paris-Dauphine, France luciano@ceremade.dauphine.fr*
[2] *Department of Statistics, Berkeley, USA, rgiordano@berkeley.edu*

Résumé. Cette présentation propose une nouvelle méthode d'inférence par simulation permettant d'estimer efficacement les postérieurs marginaux de faible dimension, sans supporter le coût computationnel d'une inférence complète en haute dimension. Notre approche exploite l'*Approximate Bayesian Computation (ABC)*, une méthode de simulation largement employée dans divers domaines, ainsi que la *Neural Ratio Estimation (NRE)*, une approche de deep learning basée sur les réseaux de neurones, afin d'améliorer leur précision et leur robustesse tout en conservant une efficacité computationnelle. Par ailleurs, nous perfectionnons la *Simulation-Based Calibration (SBC)*, un test intuitif de validité des méthodes d'inférence bayésienne, afin d'assurer une validation plus rigoureuse en pratique. Plus précisément, nous contraignons conjointement les espaces des paramètres et des données via l'ABC, permettant ainsi d'obtenir de meilleures performances à proximité des données observées. Cette approche présente un fort potentiel d'application en économie, génétique, cosmologie, épidémiologie et politiques publiques, où une inférence bayésienne fiable et scalable est essentielle.

Mots-clés. Inférence Bayésienne, Méthodes Basées sur la Simulation, Apprentissage Profond, Statistiques Computationnelles

Abstract. This presentation introduces a novel Simulation-Based Inference (SBI) method that efficiently estimates low-dimensional marginal posteriors without incurring the computational cost of full high-dimensional inference. Our approach leverages Approximate Bayesian Computation (ABC), a widely used simulation-based method across various fields, and Neural Ratio Estimation (NRE), a deep learning approach based on neural networks, to enhance accuracy and robustness while maintaining computational efficiency. Furthermore, we refine Simulation-Based Calibration (SBC), an intuitive validity test for Bayesian inference methods, to ensure more rigorous validation in practice. More specifically, we jointly constrain the parameter and data spaces through ABC, leading to improved performance in the vicinity of observed data. This method holds significant potential for applications in economics, genetics, cosmology, epidemiology, and public policy, where reliable and scalable Bayesian inference is crucial.

Keywords. Bayesian Inference, Simulation-Based Methods, Deep Learning, Computational Statistics

1 Introduction

L'inférence bayésienne constitue un cadre rigoureux permettant d'exploiter des données observées afin de mettre à jour les connaissances sur des distributions de probabilité inconnues. Elle est largement employée dans de nombreux domaines scientifiques, allant de la physique théorique aux sciences sociales et aux politiques publiques [Von Toussaint, 2011, Gelman et al., 2013], en s'appuyant sur des méthodes inférentielles similaires, quel que soit le champ d'application.

Dans une approche bayésienne, on dispose d'un ensemble de données x_{obs} supposées générées par un modèle probabiliste défini par une fonction de vraisemblance $f(x_{obs} \mid \theta)$, qui dépend de paramètres inconnus θ. Les croyances du statisticien sur ces paramètres sont exprimées à travers une distribution *a priori*, notée $\pi(\theta)$. L'inférence bayésienne consiste alors à déterminer la distribution *a posteriori*, ou *postérieure*, définie par :

$$\pi(\theta \mid x_{obs}) = \frac{\pi(\theta)\, f(x_{obs} \mid \theta)}{p(x_{obs})}, \tag{1}$$

où $p(x_{obs})$, appelé *évidence marginale*, assure la normalisation. Ce cadre permet ainsi de combiner les connaissances a priori sur les paramètres θ avec l'information fournie par les données observées x_{obs}.

En guise d'exemple, dans Heidemanns et al. [2020], les données x_{obs} représentent des sondages d'opinion réalisés avant une élection nationale aux États-Unis, tandis que les paramètres θ décrivent l'évolution temporelle des intentions de vote par État ainsi que d'autres sources de biais propres aux sondages. La distribution *a posteriori* bayésienne $\pi(\theta \mid x_{obs})$ représente l'ensemble des paramètres plausibles ayant pu générer ces observations.

Souvent, en pratique, on ne s'intéresse qu'à un faible nombre de paramètres. Dans l'exemple des sondages, l'intérêt peut porter uniquement sur l'opinion publique au niveau national le jour de l'élection. Cependant, la distribution a posteriori complète $\pi(\theta \mid x_{obs})$ englobe *l'ensemble* des paramètres inconnus, y compris de nombreux paramètres dits *nuisibles*, qui ne sont pas directement pertinents.[1]

Toutefois, ces paramètres interagissent de manière complexe avec les données, rendant généralement nécessaire l'inférence sur la loi a posteriori complète. Ce problème persiste même dans les techniques bayésiennes avancées issues de l'apprentissage automatique [Blei et al., 2017, Giordano et al., 2024]. Cette complexité rend l'estimation d'une unique marginale de la distribution postérieure bayésienne particulièrement coûteuse, voire impraticable dans certains cas.

De plus, dans de nombreuses applications pratiques, la fonction de vraisemblance $f(x_{obs} \mid \theta)$ est indisponible ou numériquement trop coûteuse à évaluer. Pour pallier cette difficulté, des méthodes d'inférence sans vraisemblance, dites *likelihood-free*, également connues sous le nom de *Simulation-Based Inference (SBI)*, ont été développées [Cranmer et al., 2020]. Ces approches reposent sur la simulation de données à partir du modèle, plutôt que sur l'évaluation directe de la vraisemblance.

Notre approche repose sur la synthèse de trois méthodes clés issues de la littérature SBI : l'*Approximate Bayesian Computation (ABC)*, l'une des premières et des mieux étudiées ; la *Neural Ratio Estimation (NRE)*, une technique moderne d'apprentissage automatique utilisant des classifieurs neuronaux ; et le *Simulation-Based Calibration (SBC)*, un outil de validation permettant d'évaluer la fiabilité des inférences bayésiennes approximatives.

Notre contribution principale repose sur l'observation que, bien qu'ABC et NRE soient toutes deux des méthodes d'inférence approximative, elles présentent des *modes d'échec opposés* et peuvent ainsi se compenser mutuellement lorsqu'elles sont combinées. Cette intégration permet une inférence plus robuste et plus efficace, moins sensible aux choix d'hyperparamètres que chacune des méthodes employée séparément. De plus, en restreignant les simulations aux régions proches des données observées, l'approche combinée améliore la capacité de SBC à identifier d'éventuelles erreurs dans l'inférence des paramètres.

2 Méthodes existantes

2.1 Approximate Bayesian Computation (ABC)

L'Approximate Bayesian Computation (ABC) [Marin et al., 2012] est l'une des premières et principales méthodes de la littérature SBI. ABC approxime la distribution a posteriori en retenant les couples paramètre-donnée (θ, x) échantillonnés à partir de la distribution jointe $\pi(\theta, x) = \pi(\theta) f(x \mid \theta)$. Ces couples sont sélectionnés en fonction de leur proximité avec les données observées x_{obs}, généralement mesurée à l'aide d'une métrique basée sur des statistiques résumées des données.

[1] Giordano et al. [2024] rapporte que, dans un ensemble de données publiques portant sur la prévision de l'élection présidentielle américaine de 2016, le modèle de sondage de Heidemanns et al. [2020] comptait plus de 15 000 paramètres nuisibles et nécessitait plus de 10 heures de calcul via Markov Chain Monte Carlo (MCMC), la méthode standard d'échantillonnage de la distribution postérieure bayésienne.

Cette méthode conduit à une distribution *pseudo-a posteriori* jointe, exprimée comme :

$$\pi_\epsilon(\theta, x \mid x_{obs}) \propto \pi(\theta) f(x \mid \theta) K_\epsilon(x_{obs} - x), \tag{2}$$

où K_ϵ est une fonction noyau (typiquement uniforme), et où le seuil ϵ contrôle la tolérance à l'acceptation des données simulées.

À partir de cette distribution jointe, sa marginale en θ est obtenue par :

$$\pi_\epsilon(\theta \mid x_{obs}) \propto \pi(\theta) \int f(x \mid \theta) K_\epsilon(x_{obs} - x)\, dx \propto \pi(\theta) g_\epsilon(x_{obs} \mid \theta), \tag{3}$$

où $g_\epsilon(x_{obs} \mid \theta)$ constitue une approximation de la fonction de vraisemblance.

Bien que conceptuellement simple et intuitif, ABC présente des limitations à la fois théoriques et pratiques. Si le seuil d'acceptation est trop large, l'approximation obtenue manque de précision ; à l'inverse, un seuil trop restrictif entraîne une explosion du nombre de simulations nécessaires, dont la complexité croît exponentiellement avec la dimension de l'espace des données. Bien que des variantes telles que Markov Chain Monte Carlo (MCMC) et Sequential Monte Carlo (SMC) aient été développées pour atténuer ces limitations, ABC reste difficile à appliquer aux problèmes de grande dimension. Néanmoins, cette méthode bénéficie d'une littérature riche offrant des bases théoriques solides ainsi qu'une mise en œuvre relativement simple, ce qui en fait une approche attrayante dans des domaines appliqués où la vraisemblance est intractable mais où l'expertise de l'utilisateur permet d'identifier des statistiques résumées pertinentes.

2.2 Neural Ratio Estimation (NRE)

La Neural Ratio Estimation (NRE) est une approche récente d'inférence par simulation exploitant l'apprentissage profond pour estimer efficacement les distributions postérieures des paramètres d'intérêt. Plutôt que de calculer directement la vraisemblance, un réseau de neurones est entraîné à distinguer les couples (θ, x) issus de la distribution jointe $\pi(\theta, x)$ de ceux obtenus indépendamment des marginales $\pi(\theta)p(x)$. Le discriminateur optimal est donné par :

$$d^*(\theta, x) = \frac{\pi(\theta, x)}{\pi(\theta, x) + \pi(\theta)p(x)},$$

permettant ainsi de définir le ratio de vraisemblance à évidence (Likelihood-to-Evidence Ratio, LTE) comme :

$$r(\theta, x) = \frac{d^*(\theta, x)}{1 - d^*(\theta, x)} = \frac{\pi(\theta, x)}{\pi(\theta)p(x)} = \frac{f(x|\theta)}{p(x)} = \frac{\pi(\theta|x)}{\pi(\theta)}.$$

Ce classifieur permet ensuite d'estimer la distribution postérieure d'intérêt par :

$$\hat{\pi}(\theta|x) = \hat{r}(\theta, x)\pi(\theta) \approx \pi(\theta \mid x),$$

et ce, pour tout jeu de données, faisant de NRE une méthode *amortie*. NRE peut également être adaptée pour estimer directement des marginales de faible dimension, ce qui est particulièrement utile lorsqu'un sous-ensemble de paramètres seulement est d'intérêt [Miller et al., 2021].

Toutefois, pour apprendre précisément la distribution de ces paramètres d'intérêt, NRE doit être entraînée sur des simulations couvrant l'ensemble des jeux de données possibles. De plus, la précision de NRE est garantie, au mieux, uniquement en moyenne sur l'ensemble des jeux de données possibles, sans garantie spécifique pour les données effectivement observées. En conséquence, nous avons constaté que même des classifieurs neuronaux de haute qualité offrent des performances inférieures à celles du MCMC sur les données réelles, et que leur efficacité diminue lorsque les distributions *a priori* sont plus diffuses ou lorsque la dimension des données augmente. Enfin, comme tous les modèles d'apprentissage profond, NRE souffre d'un manque d'interprétabilité, ce qui complique le diagnostic des erreurs et peut engendrer une propagation d'imprécisions dans l'inférence finale.

2.3 Simulation-Based Calibration (SBC)

La Simulation-Based Calibration (SBC) [Talts et al., 2020] est un outil basé sur la simulation permettant d'évaluer la qualité des inférences obtenues. Son principe est simple : si une estimation est correcte, alors les paramètres simulés à partir d'une approximation de la postérieure issue d'un jeu de données synthétique devraient suivre la distribution *a priori* :

$$\pi(\theta) = \int \pi(\tilde{\theta}) f(\tilde{x} \mid \tilde{\theta}) \pi(\theta|\tilde{x}) \, d\tilde{\theta} d\tilde{x}. \tag{4}$$

Vérifier que ces échantillons respectent bien cette distribution constitue un test formel de la précision de l'estimation.

Cependant, à l'instar de NRE, SBC repose sur une validation marginale sur l'ensemble des jeux de données possibles. Il est donc possible qu'une estimation soit très imprécise pour les données effectivement observées tout en restant correcte en moyenne, rendant cette erreur indétectable via SBC. Par ailleurs, bien que SBC puisse théoriquement être appliqué à toute approximation, sa mise en œuvre nécessite d'inférer les paramètres d'intérêt un très grand nombre de fois, ce qui le rend computationnellement prohibitif pour valider des approches classiques comme MCMC ou ABC. En revanche, il peut être facilement utilisé pour des méthodes amorties, telles que le NRE.

3 Contribution

Comme décrit précédemment, ABC et NRE présentent des limitations opposées : ABC produit une estimation précise de la postérieure lorsque les simulations sont proches des données observées, mais devient imprécis lorsque celles-ci sont trop dispersées, tandis que NRE présente un comportement inverse. De plus, SBC, applicable à NRE (une méthode dite amortie), souffre des mêmes limitations et perd en précision lorsque les simulations s'éloignent des données observées.

3.1 ABC-NRE

Notre proposition, **Neural Ratio Estimation guidé par ABC (ABC–NRE)**, exploite cette *anti-symétrie* en utilisant des techniques de simulation ABC bien connues et largement étudiées pour générer les données d'entraînement d'un classifieur NRE. Ainsi, nous considérons le modèle joint défini par (2) et ses deux marginales :

$$\pi_\epsilon(\theta \mid x_{obs}) \propto \pi(\theta) g_\epsilon(x_{obs} \mid \theta) \quad \text{et} \quad \pi_\epsilon(x \mid x_{obs}) \propto p(x) K_\epsilon(x_{obs} - x).$$

Ce faisant, nous contraignons de manière efficace et interprétable à la fois l'espace des paramètres et celui des données, garantissant ainsi que l'inférence se concentre uniquement sur les régions pertinentes par rapport aux données observées. En utilisant ces distributions ajustées, l'objectif du classificateur devient :

$$d^*_\epsilon(\theta, x) = \frac{\pi_\epsilon(\theta, x \mid x_{obs})}{\pi_\epsilon(\theta, x \mid x_{obs}) + \pi_\epsilon(\theta \mid x_{obs}) \pi_\epsilon(x \mid x_{obs})} = \frac{1}{1 + \frac{\pi_\epsilon(\theta|x_{obs})}{\pi(\theta|x)}},$$

À partir de cette formulation, on peut déduire une approximation de la distribution a posteriori :

$$\hat{\pi}(\theta \mid x) = \frac{\hat{d}_\epsilon(\theta, x)}{1 - \hat{d}_\epsilon(\theta, x)} \pi_\epsilon(\theta \mid x_{obs}).$$

Bien que le dernier terme soit typiquement intractable, si θ est de faible dimension, $\pi_\epsilon(\theta \mid x_{obs})$ peut être facilement approximée en échantillonnant depuis la distribution jointe $\pi_\epsilon(\theta, x|x_{obs})$.

Lorsqu'il est entraîné sur des données issues d'ABC, NRE agit comme une correction de l'ABC. En effet, le biais induit par NRE est directement et inversement lié à celui introduit par l'approximation ABC sur le même ensemble de simulations. Une simple combinaison formelle des approximations ABC et NRE permet d'annuler ces deux biais, produisant ainsi une estimation *invariante au seuil d'acceptation d'ABC*. Par conséquent, nous pouvons fixer un seuil d'acceptation suffisamment élevé pour garantir l'efficacité computationnelle de notre méthode.

3.2 ABC-SBC

Par ailleurs, l'avantage de la simulation par ABC bénéficie également à SBC, permettant une détection plus fine des erreurs d'estimation des paramètres qui, autrement, pourraient passer inaperçues dans une validation globale. Ainsi, SBC devient un outil pratique de validation non seulement pour ABC–NRE, mais aussi pour toute autre méthode amortie (comme les Normalizing Flows, par exemple).

4 Conclusion

En combinant l'ABC avec des méthodes telles que la NRE et la SBC, les méthodes proposées améliorent considérablement la précision de l'inférence et la qualité de la calibration. L'ABC permet de concentrer l'analyse sur les régions pertinentes des données observées, garantissant ainsi des approximations a posteriori plus précises et des diagnostics plus fiables. Ces avancées soulignent le potentiel des approches basées sur l'ABC pour relever les défis de l'inférence sans vraisemblance et ouvrent la voie à de futurs développements.

References

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. 2017. URL https://doi.org/10.1080/01621459.2017.1285773.

K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. 2020. URL https://doi.org/10.1073/pnas.1912789117.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. 2013. URL https://sites.stat.columbia.edu/gelman/book/.

R. Giordano, M. Ingram, and T. Broderick. Black box variational inference with a deterministic objective: Faster, more accurate, and even more black box. 2024. URL http://jmlr.org/papers/v25/23-1015.html.

M. Heidemanns, A. Gelman, and G. E. Morris. An updated dynamic Bayesian forecasting model for the US presidential election. 2020. URL https://doi.org/10.1162/99608f92.fc62f1e1.

J. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. 2012. URL https://doi.org/10.1007/s11222-011-9288-2.

B. K. Miller, A. Cole, P. Forré, G. Louppe, and C. Weniger. Truncated marginal neural ratio estimation. 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/01632f7b7a127233fa1188bd6c2e42e1-Abstract.html.

S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. Validating Bayesian inference algorithms with Simulation-Based Calibration, 2020. URL https://doi.org/10.48550/arXiv.1804.06788.

U. Von Toussaint. Bayesian inference in physics. 2011. URL https://doi.org/10.1103/RevModPhys.83.943.

Iterated forward scheme to construct proposals for sequential Monte Carlo algorithms

Nicolas Chopin [1] & Maud Delattre [2] & Guillaume Kon Kam King [3] & Sylvain Procope-Mamert [4]

[1] *ENSAE-CREST, Institut Polytechnique de Paris, France, nicolas.chopin@ensae.fr*
[2] *Université Paris-Saclay, INRAE, MaIAGE, France, maud.delattre@inrae.fr*
[3] *Université Paris-Saclay, INRAE, MaIAGE, France, guillaume.konkamking@inrae.fr*
[4] *Université Paris-Saclay, INRAE, MaIAGE, France, sylvain.procope-mamert@inrae.fr*

Résumée. La Sequential Monte Carlo est un ensemble puissant d'algorithmes utilisés pour échantillonner à partir d'une suite de distributions. C'est utile notamment pour l'inférence bayésienne avec différents types de modèles et d'applications à des données réelles. En particulier, lorsque nous essayons de récupérer un signal caché à partir de données produites séquentiellement avec des modèles à espace d'état caché, les propositions définies canoniquement, connues sous le nom de filtre particulaire bootstrap, se comportent rarement bien et nécessitent un travail supplémentaire pour être transformées en algorithmes d'échantillonnage efficaces. Des travaux antérieurs sur les méthodes itérées basées sur un schéma à rebours ont montré comment améliorer progressivement les propositions pour atteindre un critère d'optimalité globale, mais ils nécessitent une bonne proposition initiale et ne peuvent pas être utilisés en ligne.

Nous introduisons un schéma direct qui fait le lien entre l'optimalité locale et l'optimalité globale. Ce schéma peut être utilisé en ligne ou comme initialisation d'un schéma à rebours. Nous produisons une implémentation de ce schéma avec une performance comparable aux schémas à rebours et plus robuste à de mauvaises initialisations.

Ceci est une version courte d'un article qui sera soumis et qui inclura notamment une description de l'implémentation et des illustrations numériques sur données simulées et réelles.

Mots-clés. Sequential Monte Carlo, modèle à espace d'état caché, modèle de Feynman-Kac, filtre particulaire.

Abstract. Sequential Monte Carlo is a powerful set of algorithms used to sample from a sequence of distributions. It is useful notably for Bayesian inference with different types of models and real data applications. In particular, when we try to recover a hidden signal from sequentially produced data with state-space models, the canonically defined proposals known as the bootstrap particle filter are rarely well-behaved and need extra work to be turned into useful sampling algorithms. Previous works on iterated methods based on a backward scheme has shown how to gradually improve proposals to reach a global optimality criterion, but they require a good initial proposal and cannot be used online.

We introduce a forward scheme that bridges between local and global optimality. This scheme can be used online or as an initialization for a backward scheme. We provide an implementation of this scheme with a performance comparable to backward schemes and more robust to a lack of good initializations.

This is a short version of a longer article that will be submitted and will include a description of the implementation and numerical experiments on simulated and real data.

Keywords. Sequential Monte Carlo, state-space models, Feynman-Kac models, particle filtering.

1 Introduction

1.1 Sampling particles: Feynman-Kac formalism

We are interested in performing inference on state-space models (SSM) and in designing efficient sequential Monte Carlo algorithms. To do so we begin by introducing Feynman-Kac models, which are a useful framework to describe Sequential Monte Carlo (SMC) algorithms, a set of importance sampling algorithms using particles to sample from sequences of distributions. Then we describe their application to SSM.

A Feynman-Kac model $(M_{1:T}, G_{1:T})$ of time T is composed of a path measure: an initial distribution $M_1(dx_1)$ used to sample the initial particles and Markov kernel proposals $M_t(x_{t-1}, dx_t)$ for each time $t > 1$ used to sample new particles given the previous particle, and non-negative weight functions $G_t(x_{t-1}, x_t)$ used to assign weights to each particle.

Given a Feynman-Kac model we define the target path measure at time t:

$$\nu_t(dx_{1:t}) = \frac{1}{Z_t} G_{1:t}(x_{1:t}) M_{1:t}(dx_{1:t}), \quad Z_t = \int G_{1:t}(x_{1:t}) M_{1:t}(dx_{1:t})$$

where $x_{t:s}$ denote the sequence $(x_t, \dots, x_s)$, $G_{t:s}$ denote the product of functions from G_t to G_s, and $M_{t:s}$ denote the kernel of the distribution of $x_{t:s}$ given x_{t-1} according to Markov kernels M_t to M_s.

For a given Feynman-Kac model we define the cost-to-go (from t to s) function $H_{t \to s}$ as:

$$H_{t \to s}(x_t) = \int G_{t+1:s} dM_{t+1:s}$$

which is, up to a constant, the Radon-Nikodym derivative between the target path measure ν_s and ν_t marginal to $x_{1:t}$.

The main goal of an SMC algorithm is to sample from the full target $\nu = \nu_T$ and to estimate its integration constant $Z = Z_T$.

1.2 Applications to state-space models

Our main motivation to study Feynman-Kac models comes from state-space models (SSM), which is a class of models studying a sequence of data $y_{1:T}$ with a latent space $x_{1:T}$ following a sequential dependency.

Here we denote $\pi(y_t \mid x_t)$ the data observation model, $\pi(x_1)$ the initial distribution of the latent space and $\pi(x_t \mid x_{t-1})$ the Markov transition kernels. The canonical Feynman-Kac model $(M_{1:t}, G_{1:t})$ to infer those models is called the bootstrap particle filter (BPF) and is defined as:

$$M_t(x_{t-1}, dx_t) = \pi(dx_t \mid x_{t-1})$$
$$G_t(x_{t-1}, x_t) = \pi(y_t \mid x_t).$$

In the BPF case, targets at time t correspond to the filtering distribution: $\nu_t(dx_{1:t}) = \pi(dx_{1:t} \mid y_{1:t})$, the full target ν_T is the smoothing distribution, integration constants are likelihoods of data up to time t: $Z_t = \pi(y_{1:t})$ and cost-to-go functions correspond to the likelihood of future data given the current latent value: $H_{t\to s}(x_t) = \pi(y_{t:s} \mid x_t)$

1.3 Objectives

BPF is simple but often leads to inefficient algorithms due to particle degeneracy arising from the use of $\pi(dx_t \mid x_{t-1})$ as proposal distributions $M_t(x_{t-1}, dx_t)$. The goal of our work is to refine the BPF by proposing a way to iteratively construct better Feynman-Kac models.

More precisely, given a reference Feynman-Kac model $(M_{1:T}, G_{1:T})$, we are interested in a good estimation of the full target ν and its integration constant Z. Therefore, we need a class of Feynman-Kac models preserving the full target ν_T without necessarily keeping the same target ν_t at time $t < T$.

2 Method

2.1 Auxiliary Feynman-Kac models

We choose to use auxiliary Feynman-Kac models for the ability to express a modification of a reference Feynman-Kac model $(M_{1:T}, G_{1:T})$. This constraint on the construction of the models, is useful to express the optimal models derived later on. Auxiliary Feynman-Kac models are derived from the auxiliary particle filter. They can be defined with proposal kernels $(\bar{M}_{1:T})$ and auxiliary weights $(\eta_{0:T})$ (with $\eta_0 \equiv \eta_T \equiv 1$) such that the auxiliary Feynman-Kac model $(\bar{M}_{1:T}, \bar{G}_{1:T})$ verifies:

$$\bar{G}_t(x_{t-1}, x_t) = G_t(x_{t-1}, x_t) \frac{M_t(x_{t-1}, dx_t)}{\bar{M}_t(x_{t-1}, dx_t)} \frac{\eta_t(x_t)}{\eta_{t-1}(x_{t-1})}.$$

In this context, the target path measure $\bar{\nu}_t$ at time t verify $\bar{\nu}_t \propto \eta_t \nu_t$. Therefore, the auxiliary weight η_t changes the target at time t. The cost-to-go function $\bar{H}_{t\to T}$ is:

$$\bar{H}_{t\to T} = \int \bar{G}_{t+1:T} d\bar{M}_{t+1:T} = \frac{H_{t\to T}}{\eta_t}.$$

It can be verified that $\bar{\nu}_T = \nu_T = \nu$, so we described many Feynman-Kac models while preserving the full target. Our goal is to develop a method to construct optimal models to sample from the full target of the reference.

2.2 Estimator variance

If we denote $(x_{1:T}^{1:N}, w_{1:T}^{1:N})$ the output of an SMC algorithm with multinomial resampling and N particles described by the Feynman-Kac model from (2.1). We can define $\hat{Z}$ an estimator of the integration constant Z: $\hat{Z} = \prod_{t=1}^{T} \frac{1}{N} \sum_{n=1}^{N} w_t^n$.

This estimator is unbiased: $\mathbb{E}[\hat{Z}] = Z$ and admits the variance (Doucet and Johansen, 2011):

$$\frac{\mathbb{V}[\hat{Z}]}{Z^2} = \frac{1}{N} \sum_{t=1}^{T} \left(\int \bar{G}_t \bar{H}_{t\to T} d\nu_T \int \frac{1}{\bar{G}_t \bar{H}_{t\to T}} d\nu_T - 1 \right) + o\left(\frac{1}{N}\right). \tag{1}$$

We recall that in this context:

$$\bar{G}_t(x_{t-1}, x_t)\bar{H}_{t\to T}(x_t) = \frac{G_t(x_{t-1}, x_t) H_{t\to T}(x_t)}{\eta_{t-1}(x_{t-1})} \frac{M_t(x_{t-1}, dx_t)}{\bar{M}_t(x_{t-1}, dx_t)}.$$

The variance of this estimator has a huge impact on the difficulty to sample from the parameter space of a state space model (with particle MCMC for example). A low variance also implies a low weight degeneracy and suggests that the full proposal is not too far from the full target. That is why in the next section we use this variance to define optimality.

2.3 Optimality

We define optimality in our context as finding $\bar{M}_t$ and η_t that minimize the variance from (1). The minimizers $\bar{M}_t^*$ and η_t^* are explicit and can be written as:

$$d\bar{M}_t^* = \frac{G_t H_{t\to T} dM_t}{H_{t-1\to T}}, \quad \eta_{t-1}^* = H_{t-1\to T}.$$

The proof comes from the fact that setting (1) to 0 requires $\eta_{t-1}^* d\bar{M}_t^* \propto G_t H_{t\to T} dM_t$. This leads to an explicit backward relation:

$$\eta_t^* = H_{t\to T} = \int G_{t+1} H_{t+1\to T} dM_{t+1} = \int G_{t+1} \eta_{t+1}^* dM_{t+1} \tag{2}$$

from which we can deduce $\bar{M}_t^*$:

$$d\bar{M}_t^* = \frac{G_t \eta_t^* dM_t}{\int G_t \eta_t^* dM_t}.$$

Global optimality introduces some computational difficulties. The obvious difficulty is that this optimal is always intractable when inference for SSM is intractable and there is the need for a particle filter. This can be addressed by building approximations based on the backward scheme (2), as explored in Guarniero et al. (2017) and Heng et al. (2020). But it requires starting from the endpoint of the target path, something that is non-trivial in cases where the reference Feynman-Kac models have degenerative behaviors.

Our method uses the concept of local optimality which allows us to follow a forward scheme, which could either be used as an initial proposal for backward scheme algorithms

or be used to construct an online algorithm, something that we can't do with a backward scheme.

Here, optimizing locally is intended with the same meaning as the local optimality for a guided particle filter, described in Chopin and Papaspiliopoulos (2020). The main idea used to define this optimality is to ignore the knowledge of the full target and only include weight functions up to time t. Namely, we want to optimize the variance of the estimator $\hat{Z}_t$ of Z_t instead of $\hat{Z}_T$ of Z_T.

We look for an optimal at target path fixed (i.e. at $\eta_{1:t}$ fixed). Reducing this to minimizing the variance of the importance weights w_t^n, it is well known (Chopin and Papaspiliopoulos, 2020), that this leads to the following optimum:

$$d\bar{M}_t^* = \frac{G_t \eta_t dM_t}{\int G_t \eta_t dM_t}. \tag{3}$$

Defining local optimality gives an idea of how we can get the best proposals for a fixed path, in the next section we develop a scheme that links this local optimality to a global optimal path.

2.4 From local to global: iterated forward scheme

Since the optimal auxiliary weights are $\eta_t^* = H_{t\to T}$ we try to derive optimal models with fixed auxiliary weights $\eta_{0:T}^{(L)}$ defined as:

$$\eta_t^{(L)} = H_{t\to t+L} = \int G_{t+1:t+L} dM_{t+1:t+L}$$

that will converge to η_t^* as L increases.

We know from fixed-lag SMC literature (Doucet and Sénécal, 2004) and (Olsson et al., 2008), that the resulting target $\nu_t^{(L)}$ can approximate quickly the full target ν_T.

Those auxiliary weights verify a relation similar to the global optimal backward scheme (2):

$$\eta_t^{(L+1)} = \int G_{t+1} \eta_{t+1}^{(L)} dM_{t+1} \tag{4}$$

starting from $\eta_t^{(0)} \equiv 1$ (the reference). What is different with this scheme is that we only need to integrate up to L times to compute the auxiliary weight as opposed to T times with global optimality.

Applying local optimality (3) to fixed auxiliary weights $\eta_t^{(L)}$, we get an auxiliary model $(\bar{M}_{1:T}^{(L+1)}, \eta_{0:T}^{(L)})$ with:

$$d\bar{M}_t^{(L+1)} = \frac{G_t \eta_t^{(L)} dM_t}{\int G_t \eta_t^{(L)} dM_t}. \tag{5}$$

The computation of the integral $\eta_{t-1}^{(L+1)}$ (4) and the kernel conjugate $\bar{M}_t^{(L+1)}$ (5) both rely on the same unnormalized kernel density that we denote $u_t^{(L+1)}$. Practically, it is difficult to

learn $\eta_t^{(L)}$ constrained by the tractability of (4) and (5). However, learning $u_t^{(L)}$ simplifies implementations, since it relates to $\eta_{t-1}^{(L)}$ and $\bar{M}_t^{(L)}$ as follows

$$d\bar{M}_t^{(L)} = \frac{du_t^{(L)}}{\int du_t^{(L)}}, \quad \eta_{t-1}^{(L)} = \int du_t^{(L)}.$$

Adjusting to the training of $u_t^{(L)}$, we obtain a relation similar to (4):

$$du_t^{(L+1)} = G_t \eta_t^{(L)} dM_t = \left(G_t \int du_{t+1}^{(L)} \right) dM_t$$

from which we can learn iteratively.

3 Implementation and numerical results

We implemented our method that leverages approximations inspired by the principles of the controlled SMC algorithm (Heng et al., 2020). We compared ourselves to the same state space models, and we will present those simulated results.

The main conclusion that we draw from those results is that our method is slower since we need several steps to have a good approximation of the full-target. But in cases where the reference (or initialization) of the algorithms is bad and controlled SMC proves impossible to initialize, our method manages to recover a good approximation. This strongly suggests that our algorithm is more robust than those based on the backward scheme, and that it may be worth incurring a small computational cost to increase the likelihood of success of a given SMC method, even when used as an initialization strategy for other algorithms.

References

Chopin, N. and Papaspiliopoulos, O. (2020). *An introduction to sequential Monte Carlo*. Springer Series in Statistics. Springer, Cham.

Doucet, A. and Johansen, A. M. (2011). A tutorial on particle filtering and smoothing : Fifteen years later. In Crisan, D. and Rozovskii, B., editors, *The Oxford Handbook of Nonlinear Filtering*, pages 656–705. Oxford University Press, Oxford ; N.Y.

Doucet, A. and Sénécal, S. (2004). Fixed-lag sequential Monte Carlo. In *2004 12th European Signal Processing Conference*, pages 861–864, Vienna, Austria.

Guarniero, P., Johansen, A. M., and Lee, A. (2017). The Iterated Auxiliary Particle Filter. *Journal of the American Statistical Association*, 112(520):1636–1647.

Heng, J., Bishop, A. N., Deligiannidis, G., and Doucet, A. (2020). Controlled sequential Monte Carlo. *The Annals of Statistics*, 48(5):2904–2929.

Olsson, J., Cappé, O., Douc, R., and Moulines, É. (2008). Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli*, 14(1):155–179.

Statistique biomédicale

Potential Responses and the Order of Patient Inclusion in Early-Phase Sequential Trials.

Meliha Akouba[1] & Matthieu Clertant [2] & Alexia Iasonos [3]& John O'Quigley [4]

[1] *Laboratoire Analyse, Géométrie et Applications, Université Sorbonne Paris Nord, France, akouba@math.univ-paris13.fr*

[2] *Laboratoire Analyse, Géométrie et Applications, Université Sorbonne Paris Nord, France, clertant@math.univ-paris13.fr*

[2] *Memorial Sloan-Kettering Cancer Center, États-Unis, iasonosa@mskcc.org*

[2] *Department of Statistical Science, University College London, Royaume-Uni, john.oquigley@ucl.ac.uk*

Résumé.

Les essais cliniques de phase I sont les premières évaluations d'un nouvel agent sur des patients humains. L'objectif principal de cette phase est de déterminer la dose maximale tolérée (DMT). Différents modèles statistiques sont utilisés pour attribuer séquentiellement les patients aux différents niveaux de dose. Traditionnellement, chaque dose est associée à un taux de toxicité, défini par la proportion de patients subissant une toxicité limitant la dose (TLD). Cette approche conventionnelle suppose une large population homogène de patients, résumée par un vecteur croissant de probabilités de toxicité β, mais elle néglige l'hétérogénéité réelle de la population de l'essai, ce qui peut compromettre la précision et la reproductibilité des résultats.

Pour surmonter cela, nous considérons la population de l'essai comme fixe, ce qui signifie que nous conditionnons l'analyse aux patients spécifiquement inclus dans l'essai. Les réponses potentielles de chaque patient sont représentées sous forme d'un vecteur d'effets binaires du traitement à travers la gamme de dose, en supposant une hypothèse de monotonie. Le taux global de toxicité dans la population de l'essai est alors résumé par un vecteur croissant R. En éliminant la variabilité aléatoire de la sélection des patients dans l'essai clinique, cette approche garantit une reproductibilité totale des résultats, contrairement aux analyses basées sur des superpopulations (β). De plus, la décomposition de la variance nous permet d'évaluer l'impact de l'ordre d'inclusion des patients, offrant ainsi une évaluation plus précise des performances des méthodes.

Nos expériences montrent que, bien que les analyses traditionnelles puissent suggérer des différences mineures entre les méthodes, des écarts quantitatifs significatifs apparaissent lorsque nous utilisons R. Nous démontrons numériquement et théoriquement que les comparaisons basées sur la superpopulation β introduisent une variabilité parasite qui masque souvent les différences réelles de comportement. Par exemple, l'écart d'erreur quadratique moyenne (MSE) entre deux modèles peut sembler de 15 % sous β, mais atteint environ 25 % sous R, révélant des disparités plus marquées.

Notre méthodologie permet une évaluation approfondie des modèles basés sur des intervalles, notamment la méthode Bayesian Optimal Interval (BOIN), qui est la seule méthode

de phase I officiellement reconnue par la FDA (Food and Drug Association) comme fit-for-purpose et qui est aujourd'hui largement utilisée dans les essais oncologiques de phase précoce. Nos résultats montrent que BOIN est moins performant que les modèles dose-réponse et même que la méthode historique "3+3". De plus, sa règle d'exclusion de doses—visant à éviter la ré-évaluation de doses trop toxiques—présente une variabilité considérable en fonction de l'ordre d'inclusion des patients. Cette instabilité conduit à l'exclusion de la véritable DMT dans environ un essai sur cinq avec une taille d'échantillon standard de 36 patients, augmentant ainsi la probabilité de recommander des doses sous-optimales en termes d'efficacité.

Mots-clés. Inférence causale, Essais cliniques, Escalade de dose, Études de positionnement de dose, Comparaisons appariées, Ordre d'inclusion des patients.

Abstract.

Phase I clinical trials are the first tests of a new agent on human patients. The primary objective during this phase is to determine the maximum tolerated dose (MTD). Various statistically guided designs are used to sequentially allocate patients across dose levels. Traditionally, each dose is associated with a toxicity rate, defined by the proportion of patients experiencing dose-limiting toxicity (DLT). This conventional approach assumes a large, homogeneous patient population, summarized by an increasing vector of toxicity probability β, but overlooks the heterogeneity of the actual trial population, leading to potential issues in accuracy and reproducibility.

To address this limitation, we treat the trial population as fixed, meaning we condition on the specific patients enrolled in the trial. Each patient's potential responses are represented as a vector of binary treatment effects across dose levels, assuming monotonicity assumption. The overall toxicity rate in the trial population is then summarized by an increasing vector R. By eliminating random variability, this approach ensures full reproducibility of results, unlike analyses based on superpopulations (β). Furthermore, variance decomposition allows us to assess the impact of patient inclusion order, enabling a more precise evaluation of method performance.

Our experiments indicate that, while traditional analyses might suggest only minor differences between methods, significant quantitative differences emerge when using R. We show both numerically and theoretically that comparisons based on the superpopulation β introduce nuisance variability, which tends to obscure competitive differences. For example, the mean squared error (MSE) difference between two designs might appear to be 15% under Beta but rises to approximately 25% under R, highlighting more pronounced disparities.

Our methodology provides a comprehensive assessment of interval-based designs, including the Bayesian Optimal Interval (BOIN) method, which is the only Phase I design officially recognized as fit-for-purpose by the FDA and is now widely used in early-phase oncology trials. Our findings reveal that BOIN underperforms compared to dose-response model-based designs and even the historitical "3+3" design. Moreover, its dose exclusion rule—intended to prevent the reassessment of highly toxic doses—exhibits considerable variability depending on the order of patient inclusion. This instability results in the exclusion of the true MTD in approximately one in five trials with a standard sample size of 36 patients, ultimately increasing the probability of recommending suboptimal doses in terms of efficacy.

Keywords. Causal inference, Clinical trials, Dose escalation, Dose finding studies, Matched comparisons, Order of patient inclusion.

1 Résumé long

L'objectif global de toute méthode de positionnement de dose en oncologie est de déterminer avec précision la Dose Maximale Tolérée (DMT), en s'assurant que la plupart des patients de l'étude soient traités avec des doses ni trop élevées ni trop faibles. Identifier la DMT permet aux cliniciens d'optimiser le bénéfice pour le patient tout en minimisant les risques potentiels. Dans les modèles statistiques, la DMT est souvent définie comme le niveau de dose le plus proche d'une probabilité cible (par exemple, 25 %) de provoquer une Toxicité Limitant la Dose (TLD).

Les contraintes éthiques doivent garantir que chaque patient soit traité sur la base des meilleures informations disponibles recueillies par les chercheurs et les cliniciens. Un essai clinique de Phase I est conçu comme un processus décisionnel séquentiel, où chaque nouveau patient est affecté à un niveau de dose basé sur les observations précédentes. Divers méthodes d'escalade de dose ont été proposés pour guider l'affectation des patients et explorer soigneusement la probabilité de toxicité associée à chaque dose. L'approche basée sur un modèle implique des algorithmes de positionnement de dose qui utilisent toutes les données disponibles pour ajuster un modèle dose-toxicité. Par la suite, le patient suivant est affecté au niveau de dose ayant la probabilité de toxicité la plus proche de l'objectif préspécifié (Oquigley 1990), ou dans une approche bayésienne, à la dose la plus susceptible d'être la DMT (Clertant 2017). Une approche alternative suggère que le processus d'exploration pourrait reposer uniquement sur les observations précédentes à la dose actuelle (Ivanova et al 2007 ; Ji et al 2010 ; Liu et Yuan 2015, Cumulative Cohort Distributions, modified Toxicity Probability Interval et Bayesian Optimal Interval, respectivement). Dans ce contexte, les méthodes basés sur des intervalles doivent évaluer trois possibilités : la probabilité estimée de toxicité à la dose actuelle est trop faible, nécessitant une escalade de la dose ; elle est trop élevée, suggérant une désescalade de la dose ; ou elle est proche de l'objectif, indiquant que la dose actuelle doit être maintenue. Les méthodes basés sur des intervalles et celles basés sur un modèle, utilisent tous deux un modèle dose-toxicité pour déterminer la DMT à la fin de l'essai. Une tâche importante consiste à évaluer et à comparer les méthodes afin de choisir l'une d'elles (Azriel et al 2011).

Soit X la dose dans l'espace discret $\{1, \ldots, m\}$. Le résultat ou la réponse en termes de toxicité, Y, est 1 pour une TLD et 0 sinon. Le patient $k \in \{1, \ldots, n\}$, contribue à l'ensemble des observations avec la paire (X_k, Y_k). Un essai de positionnement de dose est un processus séquentiel dans lequel une méthode $\mathbb{M}$ recommande la dose suivante sur la base de toutes les observations précédentes :

$$X_{k+1} = \mathbb{M}(X_1^k, Y_1^k) \in \sigma(X_1^k, Y_1^k),$$

avec $\sigma(X_1^k, Y_1^k)$ l'ensemble des variables aléatoires mesurables par rapport à l'algèbre σ générée par $(X_1^k, Y_1^k) = \{(X_1, Y_1), \ldots, (X_k, Y_k)\}$. L'objectif est d'identifier et de traiter les

patients à la dose maximale tolérée (DMT) qui est la dose la plus proche en termes de probabilité de toxicité par rapport à un objectif pré-spécifié α.

Dans les études théoriques et numériques, il existe **deux principaux mécanismes aléatoires pour générer les réponses toxiques TLD.**

Le premier mécanisme suppose que les patients sont échantillonnés indépendamment d'une population homogène infinie et abstraite où, après avoir échantillonné l'individu, on lui attribue aléatoirement un label TLD en fonction d'une probabilité, cette probabilité étant obtenue à partir du niveau de dose auquel le patient est traité. La probabilité d'observer une TLD à chaque niveau de dose est donnée par le vecteur β, également appelé scénario théorique. C'est le vecteur des probabilités de toxicité à chaque dose : $\beta = (\beta_d)_{1 \leq d \leq m}$. L'hypothèse de travail habituelle pour ces essais oncologiques est que la probabilité d'une TLD augmente de manière monotone avec la dose :

$$d < d' \Rightarrow \beta_d \leq \beta_{d'}.$$

Le second mécanisme suppose une population finie incluse dans l'essai clinique. Cette population hétérogène est construite autour de $m-1$ sous-groupes, chaque sous-groupe étant identifié par le niveau de dose i, $i = 1, ..., m$, ce qui correspond au niveau de dose le plus bas où la TLD pourrait survenir. Chaque patient de la population finie a sa propre tolérance, de sorte que l'aléa est uniquement dû à l'affectation des doses aux patients. Les réponses potentielles d'un patient donné à chaque dose sont résumées par un vecteur de tolérance croissant dans $[0,1]^m$, tel que pour m niveaux de dose, il existe $m+1$ possibilités, par exemple : un patient k est associé aux données complètes $(0,0,0,1,1,1)$; cela signifie que si ce patient est traité au niveau de dose 3 ou en dessous, Y_k serait égal à 0. Le vecteur de réponse du patient i est noté $\mathbf{Y}_i = (\mathbf{Y}_{i,d})_{1 \leq d \leq m}$, où $\mathbf{Y}_{i,d}$ désigne la réponse potentielle du patient i traité à la dose d. Ainsi, pour un patient i traité à la dose X_i, le résultat observé est $Y_i = \mathbf{Y}_{i,X_i}$. Le scénario empirique, R, est le vecteur des taux de toxicité à chaque dose dans la population finie incluse dans l'essai clinique :

$$R = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i.$$

En pratique, nous n'avons que des données partielles, car la méthode assigne à chaque patient un seul niveau de dose, et nous n'observons que la réponse partielle correspondante. Cette distinction entre le vecteur complet des réponses potentielles et les observations réelles (données incomplètes) est centrale pour notre analyse, et les meilleures méthodes sont celles qui extraient et utilisent les informations les plus pertinentes à partir de ces observations partielles. Lorsque la population incluse à l'essai est fixe et qu'une méthode d'alllocation déterministe a été choisie, la seule source de variabilité restante est dûe à l'ordre d'inclusion des patients.

Le cadre causal des réponses potentielles nous aidera à évaluer la performance respective des méthodes. En particulier, nous comparerons les deux mécanismes aléatoires pour générer les réponses toxiques TLD et leur impact respectif pour comparer les méthodes. Depuis Storer 1989, les statisticiens impliqués dans les essais oncologiques de phase I utilisent le

paradigme habituel d'une grande population homogène associée au scénario théorique β pour évaluer la performance des méthodes. Nous montrons l'avantage de considérer le scénario empirique R associé à la population finie traitée durant l'essai.

Identifier la DMT est l'objectif principal des études de positionnement de dose de phase précoce. Les estimands causaux sont donc les deux DMT possibles associées soit au scénario empirique, MTD_R, soit au scénario théorique, MTD_β :

$$MTD_R = \underset{d\in\{1,\dots,m\}}{\arg\min} |\alpha - R_d| \quad \text{et} \quad MTD_\beta = \underset{d\in\{1,\dots,m\}}{\arg\min} |\alpha - \beta_d|.$$

Ce problème, fortement motivé par les idées issues de l'inférence causale (Imbens et Rubin 2015), a été exploré par des simulations approfondies. Ici, nous proposons un résultat théorique pour illustrer l'impact de l'évaluation par rapport à β ou par rapport à R.

Soit M une estimation de la probabilité de toxicité à chaque niveau de dose à la suite d'un essai clinique. On considère $\hat{M}$, $\hat{R}$ et $\hat{\beta}$, des paramètres définis sur les trois vecteurs dose/toxicité, M, R et β, respectivement. Notre paramètre d'intérêt pourrait être la DMT ou un estimateur de celle-ci, quel que soit son mode de mesure :

$$\hat{M} = MTD_M,\ \hat{R} = MTD_R \text{ et } \hat{\beta} = MTD_\beta,$$

ou une caractéristique de cet estimateur, comme la probabilité de toxicité à la DMT selon l'un des trois vecteurs dose/toxicité :

$$\hat{M} = \beta_{MTD_M},\ \hat{R} = \beta_{MTD_R} \text{ et } \hat{\beta} = \beta_{MTD_\beta}.$$

Le développement théorique suivant utilisant les notations $\hat{M}$, $\hat{R}$ et $\hat{\beta}$ s'applique à ces différentes situations. Le résultat met en évidence les différences entre les deux erreurs quadratiques.

Théorème 1: Sous des conditions générales, c'est-à-dire pour toute méthode déterministe et tout paramétre estimé, on a :

$$\mathrm{E}_\beta\left[(\hat{M}-\hat{\beta})^2\right] - \mathrm{E}_\beta\left[(\hat{M}-\hat{R})^2\right] = 2\,\mathrm{Cov}_\beta(\hat{M},\hat{R}) - \mathrm{Var}_\beta(\hat{R}) + B_m \tag{1}$$

avec

$$B_m = \left[\mathrm{E}_\beta(\hat{M}-\hat{\beta})\right]^2 - \left[\mathrm{E}_\beta(\hat{M}-\hat{R})\right]^2.$$

Nous avons montré que parmi les méthodes les plus connues en Phase I, la différence des biais B_m est faible ou négligeable. Le terme $\mathrm{Var}_\beta(\hat{R})$ est commun à toutes les méthodes car il s'agit de la variabilité de la population d'essai tirée de la super-population. Ainsi, le principal changement lorsque l'on calcule l'erreur quadratique moyenne par rapport à R plutôt que β provient de l'élimination de la covariance entre $\hat{M}$ et $\hat{R}$.

Cette covariance est un terme de pénalité pour les méthodes qui intègrent le plus efficacement l'information provenant de la population finie incluse à l'essai, résumée par R. Ces résultats sont soutenus par des simulations approfondies.

La preuve du théorème repose sur une décomposition de la variance qui nous permet de comprendre le rôle de l'ordre d'inclusion des patients. Ce concept nous sert à démontrer pourquoi les méthodes basées sur des intervalles, telle que la méthode Bayesian Optimal Interval design (BOIN), sont moins performantes que les méthodes basées sur des modèles et pourquoi les méthodes basées sur des intervalles devraient être évitées en pratique. En particulier, leur règle d'exclusion de dose — visant à prévenir la révaluation des doses trop toxiques — présente une variabilité considérable selon l'ordre d'inclusion des patients. Cette instabilité conduit à l'exclusion de la vraie DMT ou de doses plus basses dans environ un essai sur cinq avec une taille d'échantillon standard de 36 patients, augmentant finalement la probabilité de recommander des doses sous-optimales en termes d'efficacité. Ces résultats seront détaillés dans notre présentation.

Bibliographie

Azriel, D., Mandel, M. et Rinott, Y. (2011), The treatment versus experimentation dilemma in dose finding studies, *Journal of Statistical Planning and Inference*, 141(8), 2759–2768.

Clertant, M., et O'Quigley, J. (2017), Semiparametric dose finding methods, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5), 1487–1508.

Imbens, G.W. et Rubin, D.B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Ivanova, A., Flournoy, N. et Chung, Y. (2007), Cumulative cohort design for dose-finding, *Journal of Statistical Planning and Inference*, 137(7), 2316–2327.

Ji, Y., Liu, P., Li, Y. et Bekele, B. (2010), A modified toxicity probability interval method for dose-finding trials, *Clinical Trials*, 653–663.

Liu, S. et Yuan, Y. (2015), Bayesian optimal interval designs for phase I clinical trials, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(3), 507–523.

O'Quigley, J., Pepe, M. et Fisher, L. (1990), Continual reassessment method: a practical design for phase 1 clinical trials in cancer, *Biometrics*, 33–48.

Storer, B.E. (1989), Design and analysis of phase I clinical trials, *Biometrics*, 925–937.

Un cadre pour générer des données tabulaires synthétiques réalistes dans le contexte d'un essai contrôlé randomisé

(en anglais: A framework for generating realistic synthetic tabular data in a randomized controlled trial setting)

Niki Z. Petrakos [1] & Erica E.M. Moodie [2] & Nicolas Savy [3]

[1] *Université McGill, Canada, niki.petrakos@mail.mcgill.ca*
[2] *Université McGill, Canada, erica.moodie@mcgill.ca*
[3] *Université Toulouse – Jean Jaurès, France, nicolas.savy@math.univ-toulouse.fr*

Mots-clés. génération de données, IA générative, données synthétiques, données tabulaires, essais contrôlés randomisés, Generative Adversarial Network, copula

Résumé. La génération de données synthétiques réalistes a fait l'objet d'une attention particulière ces dernières années, en particulier dans le domaine de la recherche en santé. De telles données ont l'avantage, par exemple, de pouvoir être partagées tout en protégeant la vie privée des patients, ou peuvent être utilisées pour déterminer des conceptions d'essais cliniques optimaux. Alors que de nombreux travaux se sont concentrés sur la génération d'images synthétiques, la génération de données tabulaires synthétiques réalistes et complexes, du type de celles que l'on rencontre le plus souvent dans les études épidémiologiques ou cliniques classiques fait encore défaut, notamment en ce qui concerne la génération de données pour les essais contrôlés randomisés (RCT). Il n'y a pas de consensus sur la meilleure façon de générer des données tabulaires synthétiques d'essais contrôlés randomisés de manière à préserver la distribution des données multivariées sous-jacentes. Motivés par un essai contrôlé randomisé portant sur le traitement du virus de l'immunodéficience humaine, nous avons comparé empiriquement la capacité de plusieurs stratégies et de deux techniques de génération (l'une fondée sur l'apprentissage automatique, l'autre sur une méthode statistique plus classique), à reproduire fidèlement des données réalistes. Nos résultats suggèrent que l'utilisation d'une approche de génération séquentielle avec un modèle de copule R-vine pour générer des variables de base, suivi d'une simple allocation aléatoire de traitement pour imiter le cadre RCT, et des modèles de régression ultérieurs pour les variables après l'allocation de traitement (telles que le résultat de l'essai), est la manière la plus efficace de générer des données synthétiques tabulaires de RCT qui capturent des caractéristiques importantes et réalistes des données réelles.

Abstract. Generation of realistic synthetic data has garnered considerable attention in recent years, particularly in the health research domain due to its utility in, for instance, sharing data while protecting patient privacy or determining optimal clinical trial design. While much work has been concentrated on synthetic image generation, generation of realistic and complex synthetic tabular data of the type most commonly encountered in classic

epidemiological or clinical studies is still lacking, especially with regards to generating data for randomized controlled trials (RTCs). There is no consensus regarding the best way to generate synthetic tabular RCT data such that the underlying multivariate data distribution is preserved. Motivated by an RCT in the treatment of Human Immunodeficiency Virus, we empirically compared the ability of several strategies and two generation techniques (one machine learning, the other a more classical statistical method) to faithfully reproduce realistic data. Our results suggest that using a sequential generation approach with a R-vine copula model to generate baseline variables, followed by a simple random treatment allocation to mimic the RCT setting, and subsequent regression models for variables post-treatment allocation (such as the trial outcome) is the most effective way to generate synthetic tabular RCT data that capture important and realistic features of the real data.

Keywords. data generation, generative AI, synthetic data, tabular data, randomized controlled trials, Generative Adversarial Network, copula

1 Introduction

Synthetic data generation has become a topic of increased interest in many disciplines, such as finance, climate science, and health research [2, 17]. Though several data generation methods exist, many create generated data sets that are much too simplistic and cannot reflect complexities of the real world, especially with mixed data types [8, 20]. Recently in medical research, where randomized controlled trials (RCTs) are often held as the "gold standard" for determining treatment efficacy, data generation via "in silico trials" has gained attention. In silico trials are virtual trials conducted through computer simulations in order to generate synthetic data such that the generated distribution mimics the original data distribution without simply copying the original observations [15].

While machine learning (ML) methods have increased in popularity in the data generation domain, especially since the development of Generative Adversarial Networks, or GANs [10], they are not the only methods that hold promise. It has been shown that copula-based generation approaches can also work well when generating realistic synthetic data [16, 6, 22]. Copulas, and in particular R-vine copulas, are especially attractive for generating tabular RCT data because they can effectively capture univariate distributions by construction, as well as multivariate dependencies between variables. Further, there currently lacks consensus on how best to generate tabular data for RCTs. Generally, there are two ways of generating tables of data – either simultaneously and often using complex machine learning algorithms, or in a sequential fashion where a baseline population is first simulated and then population changes are simulated through predictive modeling [5, 6, 18]. We hypothesize that a sequential framework would be better suited to the task of generating tabular RCT data due to the inherently sequential nature of RCT data. We aim to empirically study methods for generating synthetic tabular RCT data by comparing the simultaneous framework to the sequential framework. Further, within the sequential framework setting, the performance of a GAN-based algorithm and a copula-based algorithm are compared.

2 Generative Algorithms

Early data generation methods, such as mean imputation or carrying-forward values from past instances, were fairly straightforward and simplistic. Though easy to implement, these methods suffered from making overly-simplistic assumptions that likely did not represent reality. We now turn to describing the methods employed in the present comparison.

2.1 GANs and CTGAN

A GAN typically consists of two fully-connected neural networks – a generative model, G, and a discriminative model, D. The generative model, $G(\boldsymbol{z}; \theta_g) : \mathcal{Z} \to \mathcal{X}$ parameterized by θ_g, takes random noise $\boldsymbol{z} \in \mathcal{Z}$ as input and outputs a vector of synthetic values $G(\boldsymbol{z}) \in \mathcal{X}$, where $\mathcal{X}$ is the real data space. The discriminative model, $D(\boldsymbol{x}; \theta_d) : \mathcal{X} \to [0, 1]$ parameterized by θ_d, takes the output from G (the vector of synthetic data) as its input, although D could of course also take real data as an input. Then, D (the discriminative model) outputs a scalar value representing the probability that its input came from the real data. Both models are trained simultaneously in an adversarial fashion, where the generative and discriminative models attempt to satisfy a minimax condition with the following value function, $V(G, D)$:

$$V(G, D) = \mathbb{E}_{x \sim p_X(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))].$$

A GAN variant designed to handle tabular data, Conditional Tabular GAN or CTGAN [21], incorporates mode-specific normalization to better mimic multi-modal continuous distributions and relies on training-by-sampling. Observations are generated by conditioning on both a discrete variable and one of the observed categories for said discrete variable, allowing for a more even exploration of all classes of each discrete variable, even for unbalanced class distributions. Since CTGANs are the preferred state-of-the-art method for generating tabular data with mixed data types [13, 4], and their implementation is well-documented through the `Synthetic Data Vault` library [16], we pursue this approach in our investigations.

2.2 Copulas and R-Vine Copula

A copula is a multidimensional cumulative distribution function (CDF) which relates the marginal distributions of a collection of random variables to the overall joint distribution. A key result that allows for the modeling of a complex joint distribution using copulas is Sklar's Theorem [19], which states that for any joint CDF, there exists a copula representation built on the marginal CDFs of each variable:

Theorem 1 (Sklar's Theorem)**.** *If* $\mathbf{X} = \{X_1, ..., X_d\}$ *is a vector of random variables with joint CDF* $F_{X_1,..,X_d}$ *and marginal CDFs* $F_{X_1}, ..., F_{X_d}$*, then there exists a copula* C *such that*

$$F_{X_1,..,X_d}(x_1, ..., x_d) = C(F_{X_1}(x_1), ..., F_{X_d}(x_d)).$$

Then, applying the chain rule and using recursive conditioning and Bayes formula, the conditional marginal distribution can be written as

$$f_{X_k|\mathbf{Z}}(x_k|\mathbf{z}) = c(F_{X_k|\mathbf{Z}_{-j}}(x_k|\mathbf{z}_{-j}), F_{Z_j|\mathbf{Z}_{-j}}(z_j|\mathbf{z}_{-j})) \times f_{X_k|\mathbf{Z}_{-j}}(x_k|\mathbf{z}_{-j}) \quad (1)$$

where $\mathbf{Z}$ is a d-dimensional vector of random variables and $\mathbf{z}$ is a vector of realizations (observations) of $\mathbf{Z}$, and subscript $-j$ indicates the entire vector excluding variable j (details of this derivation are omitted for brevity). This result shows that each conditional marginal distribution can be written as a product of bi-dimensional copulas and conditional marginal density functions, thus making copulas very useful in modeling multivariate dependencies.

Vine copulas (of which R-vine copulas are a subset) were developed as a specific type of copula that invokes a nested tree (or "vine") structure, where pairs of copulas are utilized to build the entire copula structure, and hence are useful in modeling dependencies between variables in a high-dimensional, multivariable context [12, 3, 1, 9]. Due to this utility, we chose to utilize specifically R-vine copulas as another method to consider in our framework.

3 Methods

3.1 Proposed Method

Seven generation frameworks are compared using multiple metrics and density plots. In particular, a simultaneous versus sequential approach is compared using a CTGAN. Within the sequential CTGAN approach, we also experiment with pre-processing real data that are bounded or have asymmetric distributions. Three ways of inducing randomness when generating data using sequential regression models are considered, and model complexity is examined. Within the sequential approach, R-vine copula is compared to CTGAN.

3.2 Simulation and Metrics

For the *simulataneous CTGAN* framework, a CTGAN is trained on all real data, all at once (i.e., simultaneously). This includes baseline covariates, treatment, post-randomization variables, and final outcome. Then, the trained model is used to generate observations for all variables, again simultaneously. The *sequential CTGAN* framework again involves training a CTGAN on the real data set, but data generation proceeds sequentially. The key idea is to split the real data into subsets based on the time points at which variables were collected in a trial. The first step involves only baseline variables; a CTGAN is trained on the real baseline data, and then the fitted CTGAN is used to generate observations for the baseline variables. Then, what are termed "execution models" [6, 18] are fit to each subsequent variable. In a trial setting, the next occurrence is treatment allocation, which is randomized independent of baseline information. Hence, the execution model for treatment is typically a discrete uniform distribution over the treatment options. Following this, post-randomization variables are generated via separate execution models fitted to the real data; we choose to employ statistical regression models to generate these variables.

To experiment with ways to induce randomness when generating continuous variables via regression execution models to ensure individuals with identical covariates can still have different values generated for subsequent variables, three different methods are compared: **a)** generated observation = prediction + sample from $N(0, \hat{\sigma}_{resid})$, where $\hat{\sigma}_{resid}$ is the observed standard deviation of residuals from the fitted execution model; **b)** generated observation = prediction + $\hat{e}$, where $\hat{e}$ is a randomly-sampled residual from the fitted execution model; **c)** same as **b)** but implemented by rejecting any value of prediction + $\hat{e}$ that does not satisfy the bounds of the distribution. The sequential CTGAN framework is also implemented with the first method of inducing randomness using pre-processed original data such that select variables in the real data are transformed before fitting generation models in the hopes of improving the quality of generated data. The generated data are then back-transformed so that the final synthetic data set is in the same scale as the real data. To deduce the impact of the degree of complexity of models used in the sequential framework, CTGANs are also defined for each post-treatment execution model rather than regression models. This is to determine whether fitting complex models at each stage in a sequential fit is worth the additional computational demand. When fitting CTGAN models as execution models, sequential steps are still followed such that separate models are fit for each post-randomization variable and only include variables that respect the temporal ordering.

Further, a *sequential* framework utilizing a *R-vine copula* to generate the baseline cohort is compared to the aforementioned approaches. For this framework, the execution models are identical to the regression models used in the sequential CTGAN approach, only approach **c)** is used to introduce randomness, and original data are not pre-processed. These choices were based on preliminary explorations of the methods with the CTGAN approach. See Figure 1 for a flowchart of each framework. The data generation process is repeated 500 times, and performance is compared across all simulation runs. Each run involves re-fitting models to generate a new synthetic data set and computing metrics evaluating fidelity of generated to real data. Generation via CTGANs is conducted in `python` using `Synthetic Data Vault` [16], and generation via R-vine copulas is conducted in `R` with `rvinecopulib` [14].

Univariate metrics for continuous and discrete variables include the complement of the Kolmogorov Smirnov (KS) statistic and the complement of the Total Variation Distance (TVD), respectively. As the KS statistic is the maximum distance between empirical CDFs, its complement represents distributions' "closeness". The TVD can be thought of as the categorical counterpart to the KS statistic. For bivariate comparisons, a Spearman correlation similarity score is computed for continuous variable pairs and a contingency similarity score is calculated for discrete variable pairs and mixed variable pairs. For a pair of mix-type variables, the continuous variable is dichotomized by quartiles. The correlation similarity score is defined as the normalized difference between the real and synthetic data correlation of a pair of variables. The contingency similarity score is defined as the normalized difference between the real and synthetic data proportions in a contingency table. Density plots of univariate and bivariate real and synthetic distributions are also utilized.

Similarity metrics from the ML literature, in particular precision, recall, and F-1 score, are also used [7]. First, real data are split into train and test sets, and a ML classifier is trained on the training set and used to predict the outcome values for the real data test set.

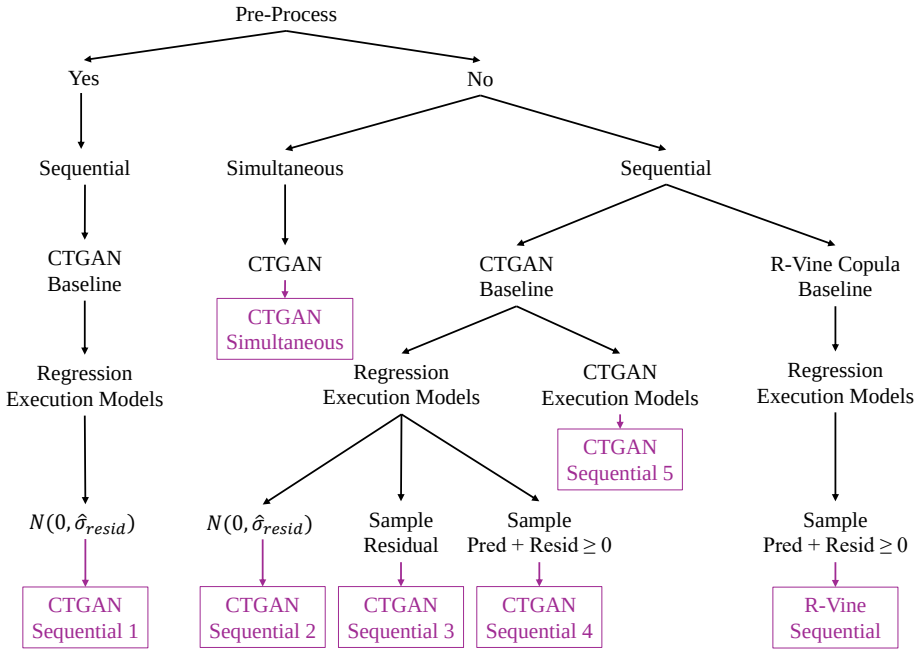


Figure 1: Flowchart of each decision point resulting in the final seven synthetic data generation frameworks.

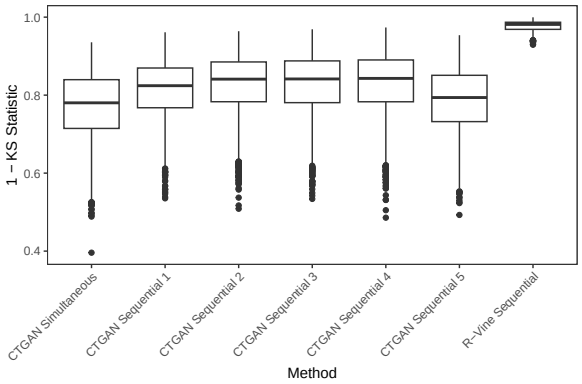
Similarly, synthetic data are split into train and test sets, and another classifier is trained on the synthetic data training set and used to predict the outcome values for the real data test set. Precision, recall, and the F-1 score are measured for each classifier. Ideally, scores from the model trained on real data and from the model trained on synthetic data should be close, thus indicating the synthetic data are similar to the real data. Since the choice of classifier may impact results, two are chosen; extreme gradient boosting and k-nearest neighbors.

3.3 Data

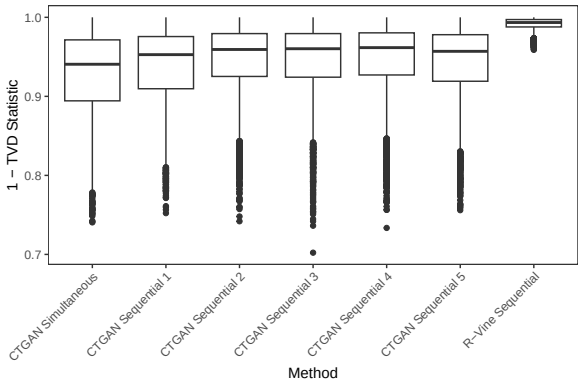
Publicly-available data from AIDS Clinical Trials Group Study 175 are used [11]. This RCT compared four human immunodeficiency virus (HIV) therapeutic arms, and 2139 people living with HIV were recruited and randomly assigned to one of four treatments. The trial outcome was whether the composite event (a CD4 cell count decline of at least 50%, progression of HIV to AIDS, or death) was observed.

3.4 Results

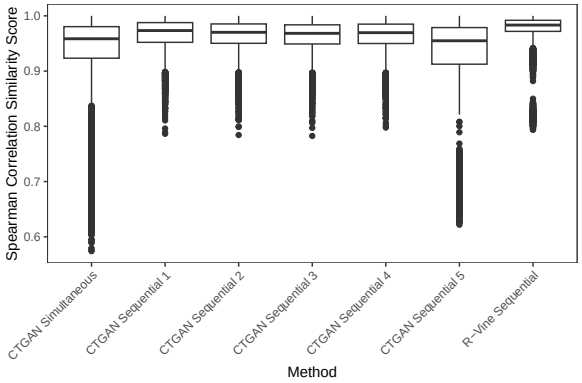
Applying comparisons of our seven frameworks using the data described above indicates sequential generation with a R-vine copula model to generate baseline variables, followed by a simple random treatment allocation to mimic the RCT setting, and subsequent regression models for variables post-treatment allocation is the most effective way to generate synthetic tabular RCT data that capture important and realistic features of the real data. Figure 2 is shown here for brevity to demonstrate these findings.



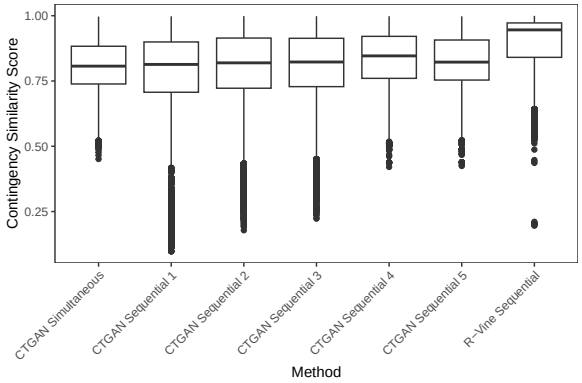
(a) Univariate Continuous Similarity Scores



(b) Univariate Discrete Similarity Scores



(c) Bivariate Continuous Similarity Scores



(d) Bivariate Discrete Similarity Scores

Figure 2: Comparison of the seven data generation methods by similarity of synthetic and real distributions; scores aggregate across variables and across all data generation runs.

4 Discussion

We recommend utilizing a vine copula approach, in particular R-vine copulas, to generate more complex tables of data (e.g., a set of baseline variables), whereas simpler regression models seem to be sufficient in generating single variables at a time to generate more realistic and meaningful synthetic data sets. Further, we recommend the use of simple statistical comparisons of univariate and bivariate distributions and careful handling of any bounded covariates to ensure faithfulness of the synthetic data to the original data on which they are based. These powerful yet simple approaches to generating realistic synthetic data can be harnessed to study design operating characteristics for new RCTs, to perform more realistic simulations when assessing finite sample performance of new methodology, or to realize "nearly" real data alongside new code without risking any breaches of data confidentiality.

References

[1] K Aas, C Czado, A Frigessi, and H Bakken. Pair-copula constructions of multiple dependence. *Insur. Math. Econ*, 44(2):182–198, 2009.

[2] S A Assefa, D Dervovic, M Mahfouz, R E Tillman, P Reddy, and M Veloso. Generating synthetic data in finance: Opportunities, challenges and pitfalls. ICAIF '20: Proceedings of the First ACM International Conference on AI in Finance, pages 1–8, New York, United States, October 15–16, 2020. Association for Computing Machinery.

[3] T Bedford and R M Cooke. Vines - A new graphical model for dependent random variables. *Ann. Stat*, 30(4):1031–1068, 1996.

[4] Q Chen, A Ye, Y Zhang, J Chen, and C Huang. An intra-class distribution-focused generative adversarial network approach for imbalanced tabular data learning. *IJMLC*, 15(7):2551–2572, 2024.

[5] J Dahmen and D Cook. SynSys: A synthetic data generation system for healthcare applications. *Sensors*, 19(5):1–11, 2019.

[6] R Demeulemeester, N Savy, P Grosclaude, N Costa, and P Saint-Pierre. Agent based modeling in health care economics: Examples in the field of thyroid cancer. *Int. J. Biostat*, 19(2):351–368, 2023.

[7] A Figueira and B Vaz. Survey on synthetic data generation, evaluation methods and GANs. *Mathematics*, 10(15):1–41, 2022.

[8] J M Franklin, S Schneeweiss, J M Polinski, and J A Rassen. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput. Stat. Data Anal*, 72:219–226, 2014.

[9] C Genest, O Okhrin, and T Bodnar. Copula modeling from Abe Sklar to the present day. *JMVA*, 201:1–9, 2024.

[10] I Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, S Ozair, A Courville, and Y Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2020.

[11] S M Hammer, D A Katzenstein, M D Hughes, H Gundacker, R T Schooley, R H Haubrich, K W Henry, M M Lederman, J P Phair, M Niu, M S Hirsch, and T C Merigan. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *NEJM*, 335(15):1081–1090, 1996.

[12] H Joe. Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. *Lecture notes-monograph series*, 28:120–141, 1996.

[13] A Kiran and S S Kumar. A methodology and an empirical analysis to determine the most suitable synthetic data generator. *IEEE Access*, 12:12209–12228, 2024.

[14] T Nagler and T Vatter. *rvinecopulib: High performance algorithms for vine copula modeling*, 02 2023. Version 0.6.3.1.1.

[15] F Pappalardo, G Russo, F M Tshinanu, and M Viceconti. In silico clinical trials: Concepts and early adoptions. *Brief Bioinform*, 20(5):1699–1708, 2019.

[16] N Patki, R Wedge, and K Veeramachaneni. The synthetic data vault. 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 399–410, Montreal, Canada, October 17–19, 2016. IEEE.

[17] V C Pezoulas, D I Zaridis, E Mylona, C Androutsos, K Apostolidis, N S Tachos, and D I Fotiadis. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *CSBJ*, 23:2892–2910, 2024.

[18] P Saint-Pierre and N Savy. Agent-based modeling in medical research, virtual baseline generator and change in patients' profile issue. *Int. J. Biostat*, 19(2):333–349, 2023.

[19] A Sklar. Fonctions de répartition à n dimmensions et leurs marges. *Annales de l'ISUP*, 8(3):229–231, 1959.

[20] Y Souli, X Trudel, A Diop, C Brisson, and D Talbot. Longitudinal plasmode algorithms to evaluate statistical methods in realistic scenarios: An illustration applied to occupational epidemiology. *BMC Med Res Methodol*, 23(242):1–15, 2023.

[21] L Xu, M Skoularidou, A Cuesta-Infante, and K Veeramachaneni. Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems, pages 1–11, Vancouver, Canada, December 8–14, 2019. NeurIPS.

[22] L B Zwep, T Guo, T Nagler, C A J Knibbe, J J Meulman, and J G C van Hasselt. Virtual patient simulation using copula modeling. *Clin. Pharmacol. Ther*, 115(4):795–804, 2024.

Score tests for non-proportional hazards in single-arm clinical trial with time-to-event endpoints

Chloé Szurewsky [1] & Guosheng Yin [2] & Gwénaël Le Teuff [1]

[1] *CESP, INSERM U1018, Université Paris-Saclay, UVSQ, Villejuif, France, chloe.szurewsky@gustaveroussy.fr, gwenael.leteuff@gustaveroussy.fr*
[2] *Department of Statistics and Actuarial Science, University of Hong Kong Pokfulam Road, Hong Kong, gyin@hku.hk*

Résumé. Sur la dernière décennie, des plans expérimentaux mono-bras en une ou deux-étapes avec un critère censuré, analogues à ceux de Simon pour un critére binaire, ont été proposés comme alternatives aux essais cliniques randomisés parfois difficiles à conduire (cancers rares, médecine personnalisée, raisons éthiques). Ces plans se basent sur le test du log-rank à un échantillon (OSLRT) ou sa version modifiée (mOSLRT) pour comparer les courbes de survie du groupe expérimental et du contrôle externe (ou historique) sous l'hypothèse des risques proportionnels (PH). Cependant, cette hypothèse peut être non vérifiée particulièrement en immuno-oncologie avec un effet retardé. Nous proposons de développer des tests du score pour les non-PH en reformulant le OSLRT de Finkelstein avec un modèle exponentiel par morceaux (PE) pour un effet précoce, central et tardif connaissant a priori le nombre et valeurs des points de changement (CPs). Pour des risques croisés, nous avons utilisé un modèle à risques accélérés. Une approche en deux étapes avec une analyse de landmark utilisant le mOSLRT est proposée pour déterminer l'effet temps-dépendant du traitement et les CPs afin de sélectionner le test du score approprié. De plus, nous avons adapté le test basé sur la moyenne des temps de survie restreints à l'horizon τ (τ-RMST) aux essais mono-bras et avons construit un test combiné (max-Combo) combinant le mOSLRT et les tests du score propres à un effet précoce et tardif. Les performances (erreur de type I, puissance) des tests developpés sont évaluées par simulation d'un essai de phase II mono-bras. Les temps de survie sont générés avec un PE pour le groupe expérimental et une distribution exponentielle pour le groupe référent (supposant aucune variabilité d'échantillonnage). Les paramètres de simulation sont la taille du groupe expérimental (20-200 patients), le taux de censure exponentielle (0-35%) et l'effet relatif du traitement (hazard ratio : 0.5-1). Comme illustration, on a utilisé un essai pédiatrique évaluant un inhibiteur pour le neuroblastome. Les simulations montrent que les tests du score sont plus conservateurs que le mOSLRT mais autant que le OSLRT. Lorsque la génération des données correspond au modèle, le test du score associé a la puissance la plus élevée même quand les CPs sont mal spécifiés. Le test basé sur le τ-RMST est aussi conservateur que le mOSLRT et moins puissant que le mOSLRT sauf pour un effet précoce avec moins de 15% de censure. Le max-Combo est conservateur, a une puissance supérieure à celle du mOSLRT pour des échantillons assez grands (n>50) mais inférieure à celle du test du score approprié pour les non-PH. Pour conclure, les tests du score sont efficients et robustes à une légère misspecification des CPs à condition que le modèle de non-PH soit connu. Inversement, le max-Combo est une alternative intéressante

lorsqu'il n'existe aucune information a priori sur la forme de l'effet traitement puisqu'il permet de combiner différents tests pour des effets du traitement PH et non-PH. Des recherches supplémentaires sont nécessaires pour étudier l'impact de la variabilité d'échantillonnage du groupe contrôle.

Mots-clés. Tests combinés, risques non-proportionnels, test du log-rank à un échantillon, modèle exponentiel par morceaux, tests du score, essais mono-bras

Abstract. In the last decade, many one or two-stage single-arm designs analog to the one or two-stage Simon's design for binary endpoints have been proposed for time-to-event endpoints as an alternative when randomised clinical trials may be difficult to conduct in rare cancers (paediatric cancers, for example), personalised medicine (biomarker-defined subsets) or considered as unethical. These designs rely on the one-sample log-rank test (OSLRT) or its modified version (mOSLRT) to compare the survival curve of an experimental arm to that of an external (historical) control group assuming the proportional hazards (PH) hypothesis. However, this hypothesis may be violated particularly when evaluating immuno-oncology clinical trials with a delayed effect of the treatment. We propose to develop statistical score tests for non-PH in reformulating Finkelstein's OSLRT score test using a piecewise exponential (PE) model for early, middle and delayed effect assuming the a priori knowledge of the number and values of change-points (CPs). For crossing hazards, we used an accelerated hazards model. A two-step approach with a landmark analysis using the mOSLRT was proposed to determine the time-dependant treatment effect and CPs to select the appropriate score test. Moreover, we adapted the restricted mean survival time at time τ (τ-RMST) based test to the case of single-arm trials and we constructed a maximum combination test (max-Combo test) that combines the mOSLRT, early and delayed score tests. The performances (type I error and power) of the developed tests are evaluated through a simulation study of a phase II single-arm trial. The survival times are generated using a PE model for the experimental group and an exponential distribution assuming no sampling variability for the reference group. The simulation parameters are the sample size of the experimental group (from 20 to 200 patients), the exponential censoring rate (from 0 to 35 %) and the relative treatment effect (hazard ratio from 0.5 to 1). For illustration, we applied these different approaches to a paediatric single-arm trial evaluating an inhibitor for neuroblastoma. Simulations show that the score tests are more conservative than the mOSLRT but as conservative as the OSLRT. When the data generation matches with the model, the associated score test has the highest power even when the CPs are misspecified. The τ-RMST-based test is as conservative as the mOSLRT and is generally less powerful than mOSLRT except for an early effect with the censoring rate less than 15%. The max-Combo test is conservative and has a higher power than the mOSLRT for enough large sample size (n>50) but less than the appropriate score test under non PH. In conclusion the developed score tests are efficient and robust to a slightly misspecification of the CP values provided that non-PH pattern is known. Conversely, the max-Combo test is an interesting alternative when no prior information exists on the pattern of treatment effect since it allows to combine different tests for PH and non PH treatment effect. Further research is needed to study the impact of the sampling variability of the external control group.

Keywords. Combination tests, non-proportional hazards, one-sample log-rank test,

piecewise exponential model, score tests, single-arm trials

1 Introduction

In oncology, well-powered phase II randomised clinical trials (RCT) with time-to-event (TTE) outcomes may be difficult to conduct in some situations such as rare cancers (paediatric cancers, for example), personalised medicine where targeted therapies are evaluated to target the particular biomarkers of a patient's tumour or for unethical reasons. The alternative is single-arm trials that compare the survival of an experimental group to that of an external (or historical) control group. In recent years, many one or two-stage designs have been proposed for single-arm trials with a TTE endpoint [Kwak and Jung (2014), Schmidt et al (2018), Wu et al (2020)]. These designs rely on the one-sample log-rank test (OSLRT) [Breslow (1975), Woolson (1981), Finkelstein et al. (2003)] and its modified version [Wu (2014)] under the proportional hazards (PH) assumption. However, this assumption can be violated when the relative treatment effect is time-dependant. This situation appears in immuno-oncology clinical trials with a delayed treatment effect, for example. The objectives of this research work are (i) to develop statistical tests adapted to single-arm trials when PH does not hold, (ii) to adapt the Restricted Mean Survival Time (RMST)-based test to single-arm trials and (iii) to evaluate other alternatives such as combination tests for single-arm trials.

2 Methods

Let T_i and C_i be the individual failure and censoring times of the i^{th} patient $(i = 1, ..., n)$ in the experimental group, that are independent and identically distributed. The observed time of event is defined as $X_i = \min(T_i, C_i)$ and the associated failure indicator as $\delta_i = I(T_i \leq C_i)$. The hazard, cumulative hazard and survival functions of the external control group are λ_0, Λ_0, S_0, respectively and those of the experimental group h_1, H_1 and S_1. Suppose that the external control group admits no sampling variability. The null and alternative hypotheses of a single-arm trial with a TTE outcome are written as follows: $H_0 : S_1(t) \leq S_0(t)$ vs $H_a : S_1(t) > S_0(t)$.
Under the PH assumption ($S_1(t) = S_0(t)^{\mathrm{HR}}$ with HR for hazard ratio), Breslow (1975), Woolson (1981) and Finkelstein et al. (2003) developed the OSLRT:

$$\mathrm{OSLRT} = \frac{\mathrm{O} - \mathrm{E}}{\sqrt{\mathrm{E}}}. \tag{1}$$

where $O = \sum_{i=1}^{n} \delta_i$ is the observed number of events and $E = \sum_{i=1}^{n} \Lambda_0(X_i)$ the expected number of events calculated from a parametric estimation of the cumulative hazard function of the external control group. However, the OSLRT is a conservative test even with large sample sizes and robust only with an exponential distribution [Wu (2014)]. So, Wu (2014) proposed

a modified version (mOSLRT):

$$\text{mOSLRT} = \frac{\text{O} - \text{E}}{\sqrt{\frac{\text{O}+\text{E}}{2}}}. \tag{2}$$

Under the null hypothesis, the OSLRT and mOSLRT follow asymptotically a standard normal distribution $\mathcal{N}(0,1)$. Hence the null hypothesis H_0 is rejected when OSLRT $< -\text{z}_{1-\alpha}$ where $z_{1-\alpha}$ is the $100(1-\alpha)$ percentile of the standard normal distribution.

2.1 OSLRT for non PH

Finkelstein et al. (2003) proposed to formulate the OSLRT as a score test defined as follows:

$$Score = \frac{\left.\frac{\partial \log(L)}{\partial \beta}\right|_{\beta=0}}{\sqrt{-\left.\frac{\partial^2 \log(L)}{\partial \beta^2}\right|_{\beta=0}}}$$

where the log-likelihood $\log(L)$ of the experimental group data is written as:

$$\log(L) \quad = \sum_{i=1}^{n} \delta_i \log(h_1(X_i)) + \log(S_1(X_i))$$

From this formulation, we developed score tests for non PH assuming a piecewise exponential (PE) survival model [Chu et al (2020)] that allows both (i) to model a time-dependant relative treatment effect between the experimental group and the external control group and (ii) to directly reformulate the $\log(L)$ of the experimental data as a function of λ_0 and S_0. Score tests with a PE model assuming an early, middle and delayed effect were also derived. This assumes that the number of changes-points (CPs) and their values are known a priori. In addition, an accelerated hazard model [Moreau et al (1992)] was also used in the setting of crossing hazards. The formulation of these four $Z-$score tests are described below:

- Early effect (EE): $Z_{\text{EE}} = \dfrac{\sum\limits_{i:X_i \le \tau} (\delta_i - \Lambda_0(X_i)) - \sum\limits_{i:X_i \ge \tau} \Lambda_0(\tau)}{\sqrt{\sum\limits_{i:X_i \le \tau} \Lambda_0(X_i) + \sum\limits_{i:X_i \ge \tau} \Lambda_0(\tau)}}$
- Middle effect (ME): $Z_{\text{ME}} = \dfrac{\sum\limits_{i:X_i \in]\tau_1;\tau_2]} (\delta_i - \Lambda_0(X_i)) + \sum\limits_{i:X_i \ge \tau_1} \Lambda_0(\tau_1) - \sum\limits_{i:X_i \ge \tau_2} \Lambda_0(\tau_2)}{\sqrt{\sum\limits_{i:X_i \in [\tau_1;\tau_2]} \Lambda_0(X_i) - \sum\limits_{i:X_i \ge \tau_1} \Lambda_0(\tau_1) + \sum\limits_{i:X_i \ge \tau_2} \Lambda_0(\tau_2)}}$
- Delayed effect (DE): $Z_{\text{DE}} = \dfrac{\sum\limits_{i:X_i > \tau} (\delta_i - \Lambda_0(X_i) + \Lambda_0(\tau))}{\sqrt{\sum\limits_{i:X_i > \tau} (\Lambda_0(X_i) - \Lambda_0(\tau))}}$
- Crossing hazards (CH): $Z_{\text{CH}} = \dfrac{\sum\limits_{i=1}^{n} \delta_i - [\Lambda_0(X_i) - \delta_i] \log(\Lambda_0(X_i))}{\sqrt{-\sum\limits_{i=1}^{n} \delta_i \log(\Lambda_0(X_i)) - [1 + \log(\Lambda_0(X_i))]\, \Lambda_0(X_i) \log(\Lambda_0(X_i))}}$

with τ the CP of an early or a delay treatment effect on the interval $[0, \tau[$ and $[\tau, \infty[$, respectively, and τ_1 and τ_2 the CPs of a middle treatment effect on the interval $[\tau_1, \tau_2]$. Under the null hypothesis, these score tests have also the same asymptotic distribution than the OSLRT i.e a standard normal distribution. As the first three Z-scores suppose the knowledge of the number and value of CPs, we propose a two-step approach with a first step consisting in a landmark analysis in order to characterise the pattern of treatment effect over time.

2.2 RMST

Restricted Mean Survival Time at time τ (τ-RMST) [Royston and Parmar (2013)] is a clinically meaningful alternative to the HR to quantify the treatment effect in RCTs for TTE outcome. One advantage of the τ-RMST is that PH hypothesis is not required. We adapted the τ-RMST-based test to single-arm trials assuming that the survival curve of the external control group admits no sampling variability:

$$\mathrm{RMST}(\tau) = \frac{\widehat{\mathrm{RMST}}_1 - \mathrm{RMST}_0}{\sqrt{\mathrm{Var}(\widehat{\mathrm{RMST}}_1)}} \tag{3}$$

where the variance of the RMST for the experimental group is estimated by the Greenwood plug-in estimator. Amongst the different ways to define the time point τ, we used that of Huang and Kuan (2018) for small sample sizes with $\tau = \min(\max(X_{i,1}), \max(X_{i,0}))$ where $X_{i,1}$ and $X_{i,0}$ are the observed survival times in the experimental and control group, respectively.

2.3 Combination tests

Combination (Combo) tests [Karrison (2016)] were developed to deal with all cases of non-proportionality in the absence of prior information on the PH or non PH patterns of the treatment effect in RCTs and are defined as the combination of weighted log-rank tests allowing to detect different types of non PH. In the setting of single-arm trials, we constructed the max-Combo test as a combination of score tests developed in Section 2.1. For example, the following test combines the mOSLRT, and score tests for early and delayed effect at different CPs:

$$\max - \mathrm{Combo} = \max(\mathrm{mOSLRT}, \mathrm{Z}_{\mathrm{EE}_{\mathrm{CP}_1}}, \mathrm{Z}_{\mathrm{EE}_{\mathrm{CP}_2}}, \mathrm{Z}_{\mathrm{DE}_{\mathrm{CP}_3}}, \mathrm{Z}_{\mathrm{DE}_{\mathrm{CP}_4}}) \tag{4}$$

Other max-Combo tests may be defined in combining different score tests with different CPs. To correct for multiple testing, we used the Hochberg correction of the p-value. We also computed the exact p-value by multiple integration as the max-Combo test asymptotically follows a multivariate normal distribution.

3 Simulation study

We performed a simulation study to evaluate the operating characteristics of a single-arm trial design with TTE endpoint based on the developed tests. Survival times of the experi-

mental group were simulated according to an exponential model for PH scenarios (scenarios 1 and 2) and a PE model for different scenarios of non PH: early effect (scenario 3), middle effect (scenario 4), delay effect (scenario 5) and crossing hazards (scenario 6). The patients of the experimental group are uniformly recruited during the first 3 years with a follow-up period of 4 years. Different CPs were a priori specified regarding this duration of the single-arm trial. The reference control group is generated with an exponential distribution with a survival median time of 2 years (λ=0.35) admitting no sampling variability. Figure 1 displays the underlying survival curves for each scenario. The simulation parameters are (i) the sample size of the experimental group $n = \{20, 30, 50, 60, 80, 100, 150, 200\}$, (ii) the exponential censoring rate $\{0, 5, 15, 25, 35\%\}$ and (iii) the treatment effect HR $= \{0.5, 0.7, 0.8, 1\}$. The type I error and power were calculated after 10 000 replications for each scenario.

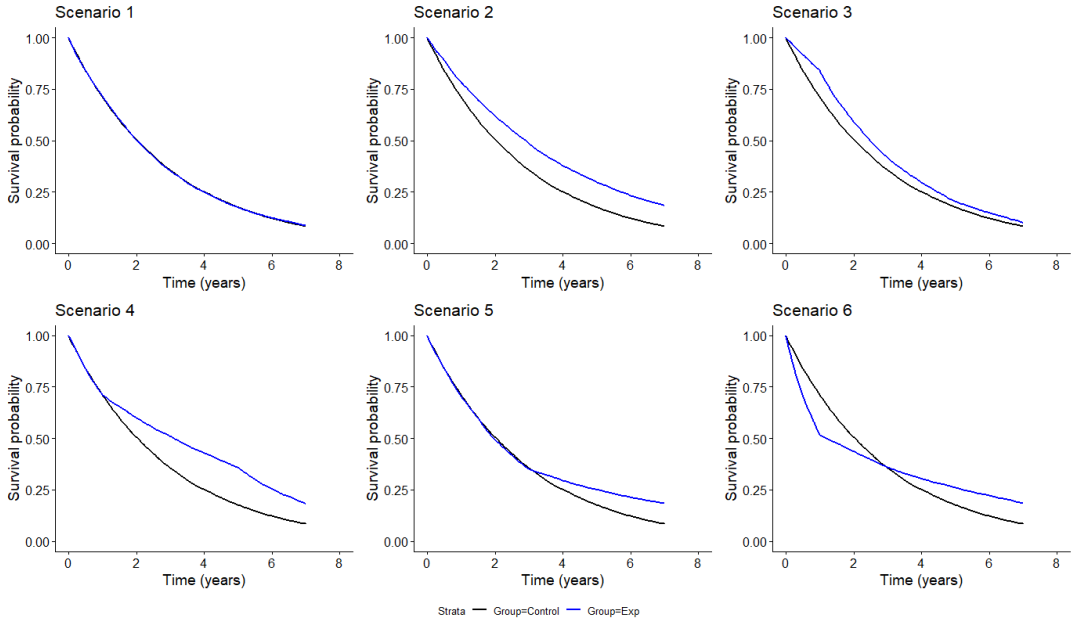


Figure 1: Theoretical overall survival curves for each scenario

In summary, the developed Z-score tests are as conservative as the OSLRT so more conservative than the mOSLRT (scenario 1). When the data generation matches with the model, the associated Z-score test is the most powerful test whatever the censoring rate and the treatment effect. When the CPs are misspecified (CP $\pm$3 months or CP $\pm$6 months), the power of the score tests decreases but remains more powerful than that of OSLRT and mOSLRT. For example, under scenario 5, Z_{DE} with misspecification of the CPs has a power reduction lower than 10%. The landmark analysis-based 2 step approach works only for $n > 100$ which limits its usefulness. Under non PH, the τ-RMST-based test is less powerful than the mOSLRT and the OSLRT except for an early effect with a censoring rate less than 15%. The max-Combo test is a conservative test (scenario 1) and is more powerful than the mOSLRT when n>50 but less than the appropriate Z-score test whatever the multiplicity correction. For example, under scenario 5, the max-Combo test has about 15% reduced power for n<100 patients and about 10% for n>100 patients compared to Z_{DE} test.

4 Illustration on real data

As an illustrative example, we used data from a phase II single-arm trial (n=91) comparing overall survival (OS) of ABT-751 (an inhibitor) to that of an external control group (n=136) in relapsed or refractory neuroblastoma children (Fox et al 2014). An early effect in favour of

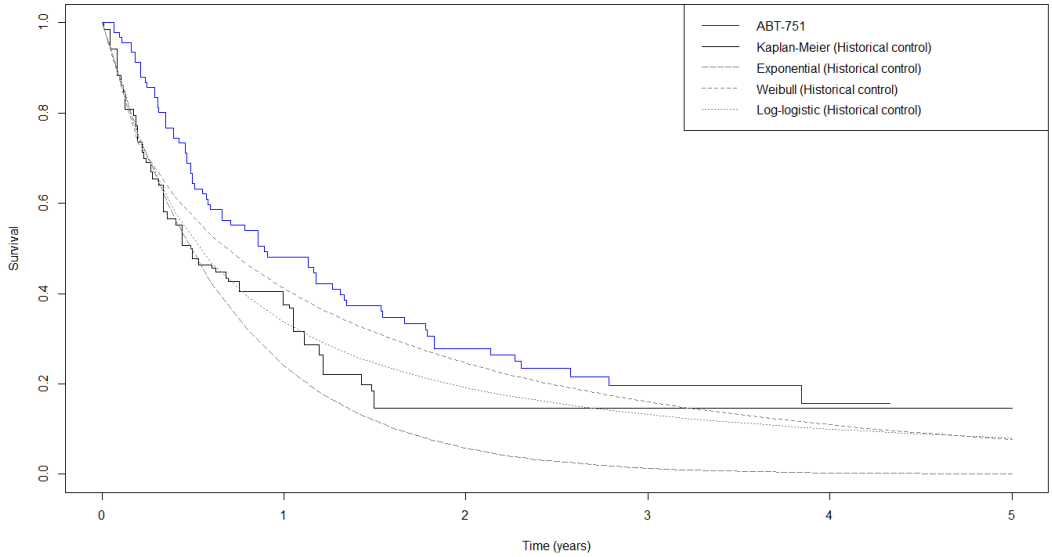


Figure 2: Kaplan-Meier curves of overall survival for the experimental group (blue line, n=91) and the external control group (black line, n=136). Dashed black lines represent different fits of external control group using parametric survival models

the experimental arm occurs in the first 2 years (Figure 2, solid lines). When an exponential distribution is used to estimate $\Lambda_0(t)$ (default distribution), all tests lead to significant results. However, the exponential distribution does not fit well the external control (highest AIC, 317) compared to Weibull (AIC = 278) and log-logistic (AIC = 244) distributions (Figure 2, dashed lines). With the log-logistic, representing the best fit, the p-value of the 2 OSLRTs, Z_{EE} tests ($\tau = 2$), τ-RMST-based test and the max-Combo test with Hochberg correction and exact p-value calculation are 0.0017, 0.0007, 0.0019, 0.0045, 0.0036 and 0.0029, respectively and a marginal effect is observed with the Z_{ME} ($\tau_1 = 0.5$ and $\tau_2 = 2$) test (p=0.0997).

5 Discussion

We proposed different score tests for one-arm oncology clinical trial designs with TTE endpoints when the PH assumption may be violated. These new score tests, constructed from PE model, are efficient under non PH and robust when the CPs values are slightly misspecified for an early, middle and late effect. We also adapted the RMST-based test to single-arm, that

has generally a smaller power than the OSLRT and the mOSLRT under non PH. We also proposed a combination test that is more powerful than the 2 OSLRTs tests for more than 50 patients but less powerful compared to the appropriate Z-score test. However, the advantage of combination tests on Z-score tests is to combine different PH and non PH treatment effect which is useful when no prior information exist on the pattern of treatment effect. Further works are needed to study the impact of the sampling variability of the external control group.

Bibliography

Breslow NE. (1975), Analysis of Survival Data under the Proportional Hazards Model, *International Statistical*, 43(1), pp. 45.

Chu C., Liu S. and Rong A. (2020), Study design of single-arm phase II immunotherapy trials with long-term survivors and random delayed treatment effect, *Pharmaceutical Statistics*, 19(4), pp. 358-369.

Finkelstein DM., Muzikansky A. and Schoenfeld DA. (2003), Comparing Survival of a Sample to That of a Standard Population, *CancerSpectrum Knowledge Environment*, 95(19), pp. 1434-1439.

Fox E., Mosse´ YP., Meany HM., et al. (2014), Time to disease progression in children with relapsed or refractory neuroblastoma treated with ABT-751: A report from the Children´ s Oncology Group (ANBL0621), *Pediatric Blood & Cancer*, 61(6), pp. 990-996.

Huang B., Kuan P.-F. (2018), Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point, *Pharmaceutical Statistics*, 17, pp. 202-213.

Kwak M. and Jung SH. (2014), Phase II clinical trials with time-to-event endpoints: optimal two-stage designs with one-sample log-rank test, *Statistics in Medicine*, 33(12).

Moreau T., Maccario J., Lellouch J. and Huber C. (1992), Weighted log rank statistics for comparing two distributions, *Biometrika*, 79(1), pp. 195-198.

Royston P. and Parmar MK. (2013), Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome, *BMC Medical Research Methodology*, 13(1), pp. 152.

Schmidt R., Faldum A., Kwiecien R. (2018), Adaptive designs for the one-sample log-rank test, *Biometrics*, 74, pp. 529-537.

Woolson RF. (1981), Rank Tests and a One-Sample Logrank Test for Comparing Observed Survival Data to a Standard Population, *Biometrics*, 37(4), pp. 687.

Wu J. (2014), A New One-Sample Log-Rank Test, *Journal of Biometrics & Biostatistics*, 05(04).

Wu J., Chen L., Wei J., Weiss H. and Chauhan A. (2020), Two-stage phase II survival trial design, *Pharmaceutical Statistics*, 19(3), pp. 214-229.

Statistique computationnelle

Classifieur Bayésien Discret Robuste en cas de Distributions Bruitées des Covariables

Wenlong Chen[1], Cyprien Gilet[1], Benjamin Quost[1] & Sebastien Destercke[1]

[1] *Université de Technologie de Compiègne, CNRS, Laboratoire Heudiasyc (Heuristics and Diagnosis of Complex Systems), CS 60 319 - 60 203 Compiègne Cedex*

Résumé. Dans cet article nous étudions le classificateur de Bayes discret (DBC), qui partitionne l'espace d'entrée en régions dans lesquelles les distributions conditionnelles par classe sont estimées. Nous étudions le partitionnement souple comme une alternative au partitionnement rigide usuel de l'espace des caractéristiques. Nous montrons que notre approche renforce la résilience du DBC face au bruit et atténue également la perte d'informations causée par ce partitionnement. Les avantages du partitionnement souple sont démontrés expérimentalement sur des bases de données réelles.

Mots-clés. Classifieur de Bayes Discret, Robustesse, Partitionnement souple.

Abstract. In this paper we study the discrete Bayes classifier (DBC), which partitions the input space into regions in which class-wise conditional probabilities are estimated. We study soft partitioning as an alternative to the usual hard partitioning of the feature space. We show that our approach enhances the resilience of DBC to noise and also mitigates the information loss caused by this partitioning. The advantages of soft partitioning are experimentally demonstrated on real-world datasets.

Keywords. Bayesian classifier, Robustness, Fuzzy partitioning.

1 Introduction

Dans cet article, nous abordons le problème de l'apprentissage d'un classificateur supervisé à partir de données bruitées. Notre objectif est d'entraîner un modèle robuste pour diagnostiquer avec précision la classe de toute nouvelle observation. Plus précisément, nous étudions comment l'assouplissement du classifieur de Bayes discret peut améliorer la robustesse face au bruit concernant les distributions des variables descriptives, ce bruit pouvant corrompre les distributions des variables descriptive concernant les données d'apprentissage ainsi que les observations de test.

Un large éventail de modèles a été proposé pour déterminer un classifieur avec des garanties de performances théoriques. Afin d'approximer au mieux la règle de décision de Bayes en présence de variables descriptives numériques ou mixtes en grandes dimensions, une approche pertinente consiste à partitionner l'espace des variables descriptives de manière à déterminer une version discrète et non paramétrique du classifieur de Bayes (DBC) (Devroye *et al.*, 1996; Gilet *et al.*, 2020). En résumé, cette approche partitionne l'espace d'entrée en régions (également appelées profils discrets), dans lesquelles le classificateur de Bayes est

déterminé analytiquement en estimant les fréquences par classe ; les instances de test étant ensuite classées en fonction des estimations dans la région à laquelle elles appartiennent. Par exemple, les profils discrets peuvent correspondre aux régions associées aux feuilles d'un arbre de classification supervisé, ou aux cellules de Voronoï dérivées d'un partitionnement Kmeans (MacQueen, 1967). La plupart des méthodes de discrétisation sont basées sur un partitionnement "rigide" : une observation de test est assignée à un profil discret unique, toutes les observations affectées à ce profil sont supposées partager des caractéristiques similaires et sont associées à la même classe prédite. John et Langley (2013) ont étudié que la discrétisation atténue l'impact du bruit ou des valeurs aberrantes. Cependant, le partitionnement "rigide" peut entraîner que des observations très similaires peuvent appartenir à des profils discrets différents auxquelles sont affectées des sorties différentes par le "DBC rigide" ultérieur. Le "DBC rigide" peut donc encore souffrir du bruit des covariables. Pour faire face à ce problème, Tang *et al.* (2002) ont étudié l'utilisation du partitionnement souple en considérant le classificateur bayésien naïf et ont obtenu des résultats prometteurs, mais pour un classificateur spécifique avec des hypothèses fortes d'indépendances, ainsi qu'une fonction de perte spécifique (0/1).

L'objectif de cet article est d'étudier l'utilisation du partitionnement souple sur le DBC, afin de bénéficier des garanties théoriques de ce dernier tout en améliorant les performances de classification issues de frontières de décision plus lisses. Nous proposons le DBC probabiliste souple (SPDBC), qui permet à chaque observation d'appartenir à différentes classes avec une certaine probabilité, permettant ainsi au modèle de gérer les données superposées ou bruyantes de manière plus flexible. Contrairement à (Tang *et al.*, 2002), notre approche permet de considérer tout type de fonction de perte/coût et sans faire d'hypothèses d'indépendances entre les variables descriptives. Des expériences sur deux bases de données réelles démontrent comment le SPDBC améliore la précision de la classification en cas de données bruitées.

2 Rappels concernant le classifieur de Bayes Discret

Contexte. Nous cherchons à calibrer un classifieur $\delta : \mathcal{X} \to \mathcal{Y}$ capable de fournir, pour toute observation $x \in \mathcal{X}$, une estimation $\delta(x)$ de sa classe réelle (inconnue) $Y \in \mathcal{Y} = \{1, \ldots, K\}$.à partir d'un ensemble d'apprentissage $\mathcal{S} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ composé d'observations $x_i \in \mathcal{X}$ appartenant à des classes $y_i \in \mathcal{Y}$. Nous considérons ici que les coûts d'erreurs peuvent ne pas être les mêmes. Définissons une matrice $L = (L_{kl})$, dont le terme général $L_{kl} \geq 0$ quantifie le coût induit par la prédiction de $\delta(x) = l$ lorsque la classe réelle est k.

Le classifieur de Bayes. La théorie de décision Bayésienne (Berger, 1985) fournit une solution théorique à ce problème d'apprentissage. Sous l'hypothèse que les données sont générées selon une probabilité conjointe $\mathbb{P}$ sur $\mathcal{X} \times \mathcal{Y}$, elle établit que la stratégie de décision minimisant le *risque attendu* (ou le coût de classification erronée) vérifie :

$$\delta^B(x) = \arg \min_{k=1,\ldots,K} R_k(\delta|x), \quad \text{avec} \quad R_l(\delta|x) = \sum_{k=1}^{K} L_{kl}\, \mathbb{P}\left(Y = k|X = x\right). \tag{1}$$

Les classifieur génératifs estiment généralement les probabilités a-posteriori $\mathbb{P}\left(Y = k|X = x\right)$ à partir des distributions conditionnelles des covariables par classe $\mathbb{P}\left(X = x|Y = k\right)$ et des probabilités a priori $\pi_k = \mathbb{P}\left(Y = k\right)$ en utilisant la formule de Bayes :

$$\mathbb{P}\left(Y = k|X = x\right) = \frac{\pi_k \mathbb{P}\left(X = x|Y = k\right)}{\mathbb{P}\left(X = x\right)}, \quad \mathbb{P}\left(X = x\right) = \sum_l \pi_l \mathbb{P}\left(X = x|Y = l\right).$$

Alors que π_k peut être estimé à l'aide des fréquences de classe dans l'ensemble d'apprentissage, l'estimation de $\mathbb{P}\left(X = x|Y = k\right)$ nécessite généralement des hypothèses supplémentaires. Ainsi, de nombreuses stratégies postulent un modèle (semi-)paramétrique pour $\mathbb{P}\left(X = x|Y = k\right)$ et se concentrent sur l'estimation des paramètres de la distribution.

Classificateur de Bayes Discret (DBC). Lorsque l'hypothèse de distribution n'est pas satisfaite, le classificateur résultant peut être une estimation biaisée du classificateur bayésien réel (John et Langley, 2013). Une alternative consiste alors à utiliser une approche discrète. Tout d'abord, l'espace d'entrée $\mathcal{X}$ est partitionné en T régions ou *profils* $\{\phi_1, \ldots, \phi_T\} = \mathcal{P}$. Nous introduisons l'application $\Phi : \mathcal{X} \to \mathcal{P}$, qui associe à toute observation x un profil $\Phi(x)$. Ensuite, le DBC revient à estimer les distributions conditionnelles $\mathbb{P}\left(X = x|Y = k\right)$ par les fractions d'échantillons d'une classe donnée appartenant aux différents profils :

$$\widehat{p}_{kt} = \widehat{\mathbb{P}}\left(\Phi(X) = \phi_t \mid Y = k\right) := \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\left\{\Phi(x_i) = \phi_t\right\}}, \tag{2}$$

avec $\mathcal{I}_k = \{i \in \{1, \ldots, n\} : Y_i = k\}$ l'ensemble des indices des observations de la classe k et $n_k = |\mathcal{I}_k|$ le cardinal de cet ensemble. Deux approximations sont faites ici : l'une est due à la discrétisation de l'espace d'entrée en profils, et donc à la substitution de la densité conditionnelle de classe $\mathbb{P}\left(X = x|Y = k\right)$ par la probabilité du profil $\mathbb{P}\left(\Phi(X) = \phi_t \mid Y = k\right)$; La seconde résulte de l'estimation de cette dernière probabilité par la fréquence relative du profil ϕ_t dans la classe k. La Figure 1 illustre un exemple 2D, où l'espace d'entrée est partitionné à l'aide de Kmeans. La DBC déterministe consiste ensuite à classer toute observation x dans la classe qui minimise le risque espéré (Gilet *et al.*, 2020) :

$$\delta_\pi^B(x) = \arg\min_{l \in \mathcal{Y}} f_l(\pi, x), \quad f_l(\pi, x) := \sum_{k \in \mathcal{Y}} L_{kl} \sum_{t=1}^{T} \frac{\pi_k \, \hat{p}_{kt}}{\sum_j \pi_j \, \hat{p}_{jt}} \mathbb{1}_{\left\{\Phi(x) = \phi_t\right\}}, \quad \forall l \in \mathcal{Y}. \tag{3}$$

3 Classifieur de Bayes Discret Souple

Dans le but de rendre le DBC moins déterministe, la Proposition suivante permet d'estimer une probabilité de prédiction à partir des quantités f_l définies en (3).

Proposition 1. *Pour toute observation $x \in \mathcal{X}$, on peut estimer la probabilité que le Classifieur de Bayes Discret δ_π^B attribue une classe $l \in \mathcal{Y}$ par*

$$\mathbb{P}\left(\delta_\pi^B(x) = l\right) = \frac{\lambda_l(\pi, x)}{\sum_{k \in \mathcal{Y}} \lambda_k(\pi, x)}, \quad \textit{où} \quad \lambda_l(\pi, x) = \sum_{q \in \mathcal{Y}} f_q(\pi, x) - f_l(\pi, x). \tag{4}$$

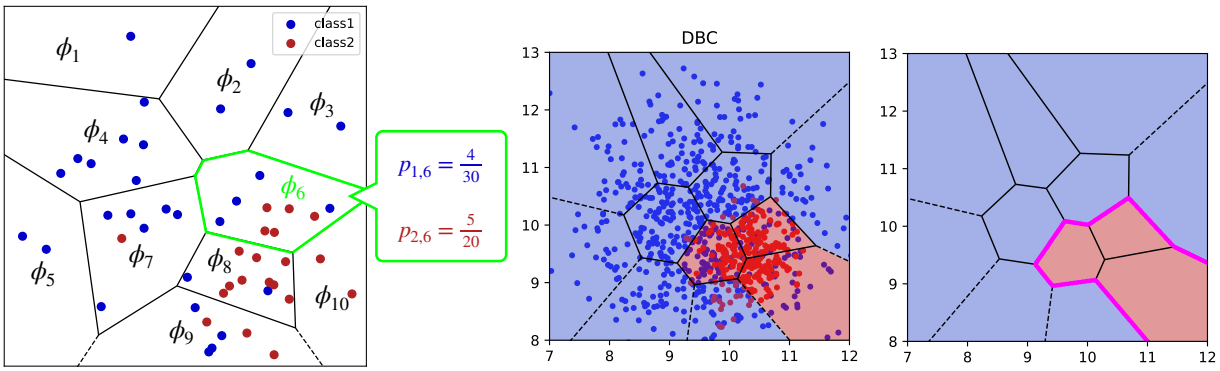


Figure 1: Problème de classification bivariée traité à l'aide du DBC. **Gauche :** Estimation des probabilités $\hat{p}_{kt}$ comme défini dans (2). **Centre et droite :** Frontières de décision associées au DBC. À chaque cellule de Voronoï est attribuée une unique classe.

Il s'en suit que la règle de décision du DBC dans l'équation (3) *équivaut à choisir la classe avec la probabilité la plus élevée* $\mathbb{P}\left(\delta_\pi^B(x) = l\right)$ *définie dans l'équation* (4) *:*

$$\delta_\pi^B : x \mapsto \underset{l \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}\left(\delta_\pi^B(x) = l\right). \tag{5}$$

Nous discutons maintenant de la manière dont les appartenances à des profils souples peuvent être exploitées pour améliorer les performances du DBC. L'idée principale est de remplacer la fonction indicatrice $\mathbb{1}_{\{\Phi(x) = \phi_t\}}$ dans l'équation (3) par une probabilité $\mathbb{P}\left(\Phi(x) = \phi_t\right)$, afin de régulariser/adoucir l'appartenance aux clusters. Cela signifie que toute méthode de clustering produisant des probabilités peut être pertinente. Dans notre cas, nous proposons d'utiliser l'algorithme Fuzzy C-means (FCM) (Bezdek *et al.*, 1984). Le choix de FCM est motivé par plusieurs raisons. Cette approche robuste, polyvalente et s'accompagne d'une interprétation géométrique ; elle fournit des représentants de cluster (les centres de cluster). Dans ce qui suit, étant donné une collection de T clusters $\{\phi_1, \ldots, \phi_T\}$ et leurs centres de cluster associés $\{c_1, ...c_T\}$, nous définissons, pour tout $x \in \mathcal{X}$ et $t \in \{1, \ldots, T\}$,

$$\mathbb{P}\left(\Phi(x) = \phi_t\right) \;=\; \frac{1}{\sum_{j=1}^{T} \left(\frac{||x - c_t||}{||x - c_j||}\right)^{\frac{2}{m-1}}}. \tag{6}$$

Définition 1. *En considérant* $\mathbb{P}\left(\Phi(x_i) = \phi_t\right)$ *définie comme dans* (6)*, la probabilité estimée qu'une observation possède le profil discret* ϕ_t *sachant que sa classe réelle est* $y = k$*, est donnée par*

$$\hat{p}_{kt} = \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \mathbb{P}\left(\Phi(x_i) = \phi_t\right). \tag{7}$$

Définition 2. *À partir de la Définition 1, le risque* f_l *de prédire la classe* $l \in \mathcal{Y}$ *pour une observation* $x \in \mathcal{X}$ *[risque initialement définie par l'équation* (3)*] devient*

$$f_l(\pi, x) := \sum_{k \in \mathcal{Y}} L_{kl} \sum_{t=1}^{T} \frac{\pi_k\, \hat{p}_{kt}}{\sum_j \pi_j\, \hat{p}_{jt}}\, \mathbb{P}\left(\Phi(x) = \phi_t\right). \tag{8}$$

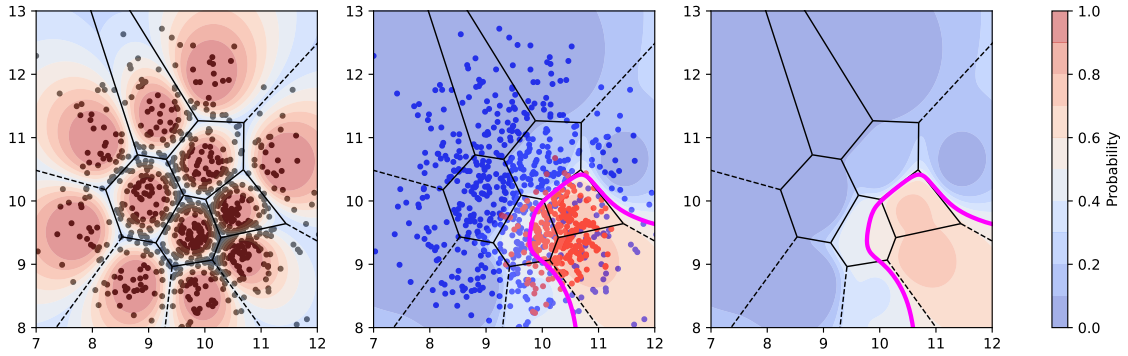


Figure 2: DBC probabiliste souple obtenue à l'aide de FCM. **Gauche.** Probabilité que chaque position appartienne au profil actuel. **Centre et droite.** Probabilité que chaque position dans l'espace des caractéristiques appartienne à la classe rouge.

Contrairement à ce que nous avons observé en Figure 1, chaque observation bruitée peut ainsi appartenir à différents profils avec des probabilités spécifiques, comme on peut le voir sur la Figure 2. Ceci permet au modèle de prendre en compte les informations voisines dans différents profils et de fournir une frontière de décision plus souple qui n'est pas limitée par les limites des clusters. Cela atténuera l'impact d'un partitionnement brut et se traduira par de meilleures performances de généralisation.

4 Expériences

L'objectif de notre étude expérimentale est de valider la robustesse et la précision de notre modèle face au bruit en comparaison avec le DBC usuel. Nous utilisons deux ensembles de données disponibles sur UCI Machine Learning : "*Diabetes*" et "*Heart Disease*" détaillés dans la Table 1. Dans le but de tester la robustesse du modèle face au bruit des variables descriptives, nous ajoutons du bruit dans chaque variable X_j dans l'ensemble de test selon une distribution gaussienne : $X_j^{\text{noise}} = X_j + \epsilon_j$ avec $\epsilon_j \sim \mathcal{N}\left(0, \sigma_j^2\right)$, $\sigma_j = \text{std}(\text{X}_\text{j}) \times s$, avec $s \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

Nous utilisons les Kmeans comme approche de clustering "rigide" et le Fuzzy C-means pour le partitionnement "souple". Afin d'obtenir une comparaison juste et indépendante du positionnement des centroïdes, nous définissons les centres de cluster dans FCM comme étant les mêmes que pour Kmeans, ce qui signifie que les centroïdes des deux approches seront exactement les mêmes, sauf que C-means permet aux données d'appartenir à différents profils avec des degrés d'appartenance spécifiques, calculés à l'aide de l'équation (6). Les résultats de cette expérience sont illustrés dans la Figure 3 et montrent que notre modèle est effectivement plus robuste que le DBC usuel en cas de données bruitées sur les observations test.

Dataset	nb. obs	nb. classes	Proportions classe	nb. variables
Diabetes	768	2	[0.65, 0.35]	8
Heart Disease	303	2	[0.54, 0.46]	13

Table 1: Informations sur les bases de données utilisées.

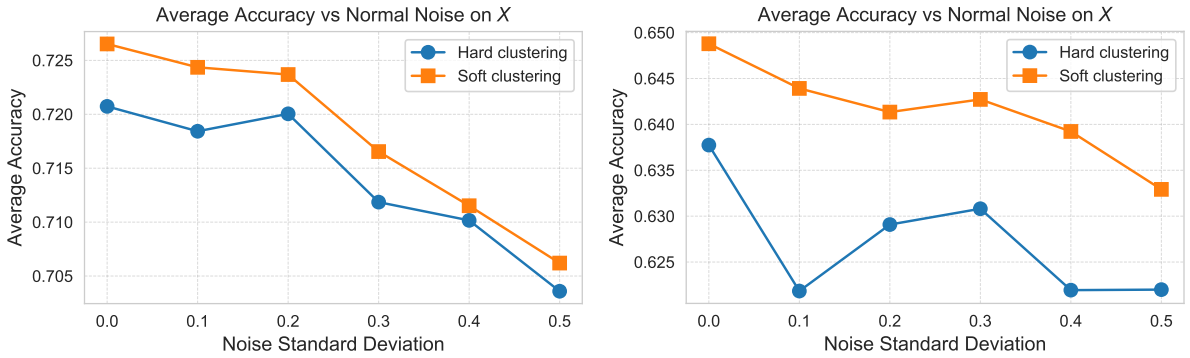


Figure 3: Résultats de l'expérience. **Gauche.** Résultats sur la base de données "*Diabetes*". **Droite.** Résultats sur la base de données "*Heart Disease*"

5 Conclusion

Dans cet article, nous avons exploré les capacités de la classification Bayésienne discrète et son adaptation aux techniques de partitionnement souple. Notre approche allie l'efficacité de la classification Bayésienne discrète à la flexibilité des profils probabilistes, améliorant ainsi la robustesse du classifieur de Bayes Discret dans des scénarios bruités.

Bibliographie

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis; 2nd ed.* Springer Series in Statistics. Springer, New York.

Bezdek, J. C., Ehrlich, R. et Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, 10(2-3):191–203.

Devroye, L., Györfi, L. et Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, 2nd édition.

Gilet, C., Barbosa, S. et Fillatre, L. (2020). Discrete box-constrained minimax classifier for uncertain and imbalanced class proportions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2923–2937.

John, G. H. et Langley, P. (2013). Estimating continuous distributions in bayesian classifiers. *arXiv preprint arXiv:1302.4964*.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

Tang, Y., Pan, W., Li, H. et Xu, Y. (2002). Fuzzy naive bayes classifier based on fuzzy clustering. *In IEEE International Conference on Systems, Man and Cybernetics*.

Un Nouvel Algorithme de Type Full AdaGrad

Antoine Godichon-Baggioni [1] & Wei Lu [2] & Bruno Portier[3]

[1] *Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, France, antoine.godichon_baggioni@upmc.fr*
[2] *MAP5, Université Paris Cité, weilu@math.cnrs.fr*
[3] *LMI UR 3226, INSA Rouen Normandie, Normandie Univ, France, bruno.portier@insa-rouen.fr*

Résumé. Une nouvelle approche est proposée pour surmonter les défis computationnels de l'algorithme Full AdaGrad pour résoudre un problème d'optimisation stochastique. En développant une méthode récursive permettant d'estimer l'inverse de la racine carrée de la covariance du gradient, cette étude propose des algorithmes efficaces et pratiques adaptés aux applications à grande échelle. Cette méthode réduit la complexité de calcul par rapport aux approches existantes. De plus, les taux de convergence des estimateurs proposés ainsi que leur efficacité asymptotique sont présentés. Leur efficacité est démontrée à travers des études numériques.

Mots-clés. Optimisation stochastique, estimation en ligne, algorithme de Robbins-Monro

Abstract. A new approach is proposed to overcome the computational challenges of the Full AdaGrad algorithm in the context of stochastic optimization. By developing a recursive method to estimate the inverse square root of the gradient covariance, this study introduces efficient and practical algorithms suitable for large-scale applications. This method reduces the computational complexity compared to existing approaches. Furthermore, the convergence rates of the proposed estimators and their asymptotic efficiency are presented. Their effectiveness is demonstrated through numerical studies.

Keywords. Stochastic Optimization, online estimation, Robbins-Monro algorithm

1 Introduction

Soit $F : \mathbb{R}^d \longrightarrow \mathbb{R}$ la fonction strictement convexe défini pour tout $\theta \in \mathbb{R}^d$ par :

$$F(\theta) := \mathbb{E}\left[f(X, \theta)\right],$$

où X est une variable aléatoire prenant ses valeurs dans un espace mesurable $\mathcal{X}$ et $f : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}$. Soit $\theta^* \in \mathbb{R}^d$ l'unique minimiseur de la fonction F que l'on suppose inconnu. On s'intéresse à l'estimation de θ^* à partir d'une suite de variables aléatoires $X_1, \cdots, X_t, \cdots$ supposées indépendantes et identiquement distribuées selon la même loi que X.

Une approche largement utilisée pour résoudre ce problème d'optimisation est l'algorithme du gradient stochastique, introduite par Robbins et Monro (1951). Malgré son efficacité computationnelle et ses bonnes propriétés de convergence, l'algorithme du gradient stochastique présente certaines limitations (Ruder, 2016). Une variante populaire de cette méthode est l'algorithme Adagrad (Adaptive Gradient) introduit par Duchi *et al.* (2011). Plus précisément, si on note pour tout $\theta \in \mathbb{R}^d$, $g_t(\theta) := \nabla_\theta f(X_t, \theta)$, où $\nabla_\theta f$ désigne le gradient de la fonction f par rapport à θ. L'algorithme Full Adagrad est défini pour tout $t \geq 1$ par :

$$\theta_t = \theta_{t-1} - \nu_t \mathcal{G}_t^{-1/2} g_t(\theta_{t-1}),$$

où θ_0 est choisi arbitrairement, $(\nu_t)_{t\geq 1}$ est une suite de pas positifs décroissants vers 0, $\mathcal{G}_t := \sum_{k=1}^{t} g_k(\theta_{k-1}) g_k(\theta_{k-1})^T$ est une estimation de la matrice de covariance du gradient, et $\mathcal{G}_t^{-1/2}$ désigne l'inverse de la racine carrée de cette matrice. Cependant, le calcul de l'inverse de cette racine carrée requiert une complexité en $\mathcal{O}(d^3)$. Pour résoudre ce problème, une version diagonale de l'Adagrad a été proposée, simplifiant le calcul en utilisant uniquement les éléments diagonaux de $\mathcal{G}_t$, soit :

$$\theta_t = \theta_{t-1} - \nu_t \mathrm{diag}\left(\mathcal{G}_t\right)^{-1/2} g_t(\theta_{t-1}). \tag{1}$$

Cet algorithme est beaucoup utilisé en pratique et se révèle très souvent plus efficace (voir par exemple (Dean *et al.*, 2012; Seide *et al.*, 2014; Smith, 2017).) Bien que plus simple à mettre en oeuvre, la version diagonale d'Adagrad perd nécessairement de l'information par rapport à la version complète.

Notre travail se concentre sur la version complète d'Adagrad, en proposant une méthode récursive pour estimer l'inverse de la racine carrée de la matrice de covariance $\Sigma := \mathbb{E}\left[\nabla_\theta f(X, \theta^*) \nabla_\theta f(X, \theta^*)^T\right]$, où θ^* minimise la fonction F. Contrairement à la version de Full Adagrad d'origine, qui utilise G_t pour estimer Σ, puis calcule $G_t^{-1/2}$, nous proposons d'estimer directement $\Sigma^{-1/2}$. En utilisant le fait que :

$$\Sigma^{-1/2}\Sigma\Sigma^{-1/2} - I_d = \mathbb{E}\left[\Sigma^{-1/2}\nabla_\theta f(X, \theta^*)\nabla_\theta f(X, \theta^*)^T\Sigma^{-1/2} - I_d\right] = 0,$$

nous introduisons un algorithme de Robbins-Monro pour estimer $\Sigma^{-1/2}$. Cet estimateur, noté A_t, est défini récursivement pour tout $t \geq 1$ par :

$$A_t = A_{t-1} - \gamma_t \left(A_{t-1} g_t(\theta_{t-1}) g_t(\theta_{t-1})^T A_{t-1} - I_d\right),$$

où $A_0 = I_d$ et $(\gamma_t)_{t\geq 1}$ est une suite de nombres réels positifs décroissants vers 0. Cet estimateur est ensuite utilisé pour mettre à jour l'estimation de θ^* :

$$\theta_t = \theta_{t-1} - \nu_t A_{t-1} g_t(\theta_{t-1}).$$

Cette approche permet d'éviter le coût computationnel élevé associé au calcul de l'inverse de la racine carrée de $\mathcal{G}_t$. Néanmoins, A_t n'est pas nécessairement défini positive, ce qui nous conduit à considérer dans ce travail, une légère modification de cet algorithme.

2 Hypothèses

On présente dans cette section les hypothèses relatives aux fonctions F et f définies dans l'introduction.

Hypothèse 1 *La fonction F est strictement convexe, deux fois continûment différentiable, et il existe $\theta^* \in \mathbb{R}^d$ tel que $\nabla F(\theta^*) = 0$.*

Hypothèse 2 *Il existe un entier $p \geq 1$ et une constante positive $C_p < \infty$ tels que, pour tout $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\|\nabla_\theta f(X,\theta)\|^{2p}\right] \leq C_p + C_p\left(F(\theta) - F(\theta^*)\right)^p.$$

Hypothèse 3 *La fonction $\Sigma : \theta \longmapsto \mathbb{E}\left[\nabla_\theta f(X,\theta)\nabla_\theta f(X,\theta)^T\right]$ est L_Σ-Lipschitzienne et $\Sigma(\theta^*)$ est définie positive.*

Cette hypothèse garantit la convergence des estimations de la variance, et plus précisément, de la racine carrée de leur inverse. Cette hypothèse est assez courante dans la littérature, en particulier lorsqu'il s'agit de l'estimation de la covariance asymptotique (Zhu *et al.*, 2023; Godichon-Baggioni et Lu, 2024).

Hypothèse 4 *Le Hessiennne de F est uniformément borné par $L_{\nabla F}$.*

Hypothèse 5 *Le Hessienne de F est localement Lipschitzien : il existe $\eta > 0$ et $L_\eta > 0$ tels que, pour tout $\theta \in \mathcal{B}(\theta^*, \eta)$,*

$$\left\|\nabla F(\theta) - \nabla^2 F(\theta^*)(\theta - \theta^*)\right\| \leq L_\eta \|\theta - \theta^*\|^2.$$

Ces hypothèses sont proches de celles rencontrées dans la littérature, notamment lorsqu'on considère l'estimation de la covariance asymptotique (Pelletier, 2000; Gadat et Panloup, 2023; Boyer et Godichon-Baggioni, 2023; Zhu *et al.*, 2023).

3 Algorithme Full AdaGrad avec une complexité en $\mathcal{O}(td^2)$

Nous nous intéressons d'abord à l'estimation récursives de la matrice $\Sigma^{-1/2}$. Rappelons que l'algorithme de Robbins-Monro pour estimer $\Sigma^{-1/2}$, décrit dans l'introduction, est défini récursivement pour tout $t \geq 0$ par

$$A_{t+1} = A_t - \gamma_{t+1}\left(A_t g_{t+1}(\theta_t) g_{t+1}(\theta_t)^T A_t - I_d\right),$$

où A_0 est une matrice symétrique définie positive choisie arbitrairement, $(\theta_t)_{t\geq 0}$ est une suite d'estimateurs de θ^*, et $\gamma_t = c_\gamma t^{-\gamma}$ avec $c_\gamma > 0$ et $1/2 < \gamma < 1$. On remarque que $A_{t+1} g_t(\theta_t)$

est un vecteur de $\mathbb{R}^d$, ce qui implique que la complexité de cette opération est en $\mathcal{O}(d^2)$. Ainsi, la complexité de la mise à jour de A_{t+1} est elle aussi en $\mathcal{O}(d^2)$. Toutefois, nous ne pouvons pas garantir que la matrice A_t soit toujours définie positive. Pour résoudre ce problème, nous proposons une version légèrement modifiée de A_t, définie pour tout $t \geq 0$ par

$$A_{t+1} = A_t - \gamma_{t+1}\left(A_t g_{t+1}\left(\theta_t\right) g_{t+1}\left(\theta_t\right)^T A_t - Id\right)\mathbf{1}_{\left\{g_{t+1}(\theta_t)^T A_t g_{t+1}(\theta_t) \leq \beta_{t+1}\right\}},$$

où $\beta_t = c_\beta t^\beta$ avec $0 < \beta < 1/2$ et $0 < c_\beta c_\gamma < 1$. Grâce à cette modification, nous pouvons montrer que A_t est définie positive pour tout $t \geq 0$. Nous pouvons maintenant proposer un algorithme Full AdaGrad défini pour tout $t \geq 0$ par

$$\theta_{t+1} = \theta_t - \nu_{t+1} A_t g_{t+1}\left(\theta_t\right), \tag{2}$$

$$A_{t+1} = A_t - \gamma_{t+1}\left(A_t g_{t+1}\left(\theta_t\right) g_{t+1}\left(\theta_t\right)^T A_t - Id\right)\mathbf{1}_{\left\{g_{t+1}(\theta_t)^T A_t g_{t+1}(\theta_t) \leq \beta_{t+1}\right\}}, \tag{3}$$

où θ_0 est choisi arbitrairement. Pour assurer l'optimalité asymptotique des estimateurs, nous introduisons l'algorithme Averaged Full AdaGrad (AFA) défini récursivement pour tout $t \geq 0$ par

$$\theta_{t+1} = \theta_t - \nu_{t+1} A_t g_{t+1}\left(\theta_t\right) \tag{4}$$

$$\overline{\theta}_{t+1} = \left(1 - \frac{1}{n+1}\right)\overline{\theta}_{t+1} + \frac{1}{n+1}\theta_{t+1} \tag{5}$$

$$A_{t+1} = A_t - \gamma_{t+1}\left(A_t g_{t+1}\left(\overline{\theta}_t\right) g_{t+1}\left(\overline{\theta}_t\right)^T A_t - Id\right)\mathbf{1}_{\left\{g_{t+1}\left(\overline{\theta}_t\right)^T A_t g_{t+1}\left(\overline{\theta}_t\right) \leq \beta_{t+1}\right\}} \tag{6}$$

$$\overline{A}_{t+1} = \left(1 - \frac{1}{n+1}\right)\overline{A}_{t+1} + \frac{1}{n+1}A_{t+1} \tag{7}$$

avec $\overline{\theta}_0 = \theta_0$, $\overline{A}_0 = A_0$.

Dans le théorème suivant, nous établissons la consistance forte des estimateurs de Σ^{-1} ainsi que les vitesses de convergence presque sûre des estimateurs de θ^*.

Theorem 3.1 *Supposons que les hypothèses 1, 3, 4 et 5 sont vérifiées, ainsi que 2 avec* $p > \max\left\{\frac{8-8\gamma}{\gamma+\beta-1}, 2\left(\frac{1}{\gamma}-1\right)\right\}$. *Supposons également que* $2\gamma + 2\nu > 3$, $\nu + \beta < 1$, $2\gamma - 2\beta > 1$ *et* $\gamma + \beta > 1$. *Alors,*

$$A_t \xrightarrow[t\to+\infty]{a.s} \Sigma^{-1/2} \qquad et \qquad \overline{A}_t \xrightarrow[t\to+\infty]{a.s} \Sigma^{-1/2}.$$

De plus,

$$\left\|\overline{\theta}_t - \theta^*\right\|^2 = O\left(\frac{\ln t}{t}\right) \quad a.s. \qquad et \qquad \sqrt{t}\left(\overline{\theta}_t - \theta^*\right) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right)$$

avec $\Sigma := \Sigma(\theta^*)$ *et* $H := \nabla^2 F(\theta^*)$.

Ainsi, nous obtenons l'efficacité asymptotique de l'estimateur moyenné. De plus, ces derniers ne nécessitent que $\mathcal{O}(td^2)$ opérations, alors que le calcul direct de $\mathcal{G}_t^{-1/2}$ impliquerait une complexité de l'ordre de $\mathcal{O}(td^3)$.

4 Expérimentations numériques

Nous comparons les performances de Full AdaGrad, AdaGrad et de leurs versions moyennées, en considérant le modèle de régression linéaire. Soit (X, Y) un vecteur aléatoire prenant ses valeurs dans $\mathbb{R}^d \times \mathbb{R}$. Considérons le cas où X est un vecteur aléatoire gaussien centré, et où

$$Y = X^T \theta^* + \varepsilon,$$

où θ^* est un paramètre de $\mathbb{R}^d$, et $\varepsilon \sim \mathcal{N}(0, 1)$ est indépendant de X. Si la matrice $\mathbb{E}[XX^T]$ est définie positive, alors θ^* est l'unique minimiseur de la fonction F définie pour tout $h \in \mathbb{R}^p$ par

$$F(h) = \frac{1}{2}\mathbb{E}\left[(Y - h^T X)^2\right].$$

Dans les simulations suivantes, nous fixons $d = 20$. Pour chaque échantillon, nous simulons $N = 30000$ réalisations de $X \sim \mathcal{N}(0, \Sigma_X)$, où Σ_X est une matrice de covariance définie positive. Notons que, dans ce cas, la variance du gradient satisfait $\Sigma = \Sigma_X$. Nous considérons deux structures différentes pour Σ_X. La première est $\Sigma_X = I_d$, ce qui correspond au cas de prédicteurs indépendants. La seconde est $\Sigma_X = R$ avec $R_{i,j} = 0.9^{|i-j|}$, ce qui induit une forte corrélation entre les prédicteurs. Pour comparer les deux algorithmes, nous calculons l'erreur quadratique moyenne de la distance entre θ_t et θ^* en moyennant sur 100 échantillons.

Comme le montre la Figure 1, lorsque Σ_X est la matrice identité, AdaGrad et Full AdaGrad ont des performances quasi identiques. Dans ce cas, Σ est une matrice diagonale. Ainsi, lorsque AdaGrad n'exploite que les éléments diagonaux, il ne perd aucune information. Cependant, lorsque les prédicteurs sont fortement corrélés et que les éléments hors-diagonale de Σ ne sont plus nuls, Full AdaGrad surpasse nettement AdaGrad. Cela souligne l'importance d'utiliser Full AdaGrad plutôt que AdaGrad lorsque la matrice de variance n'est pas diagonale.

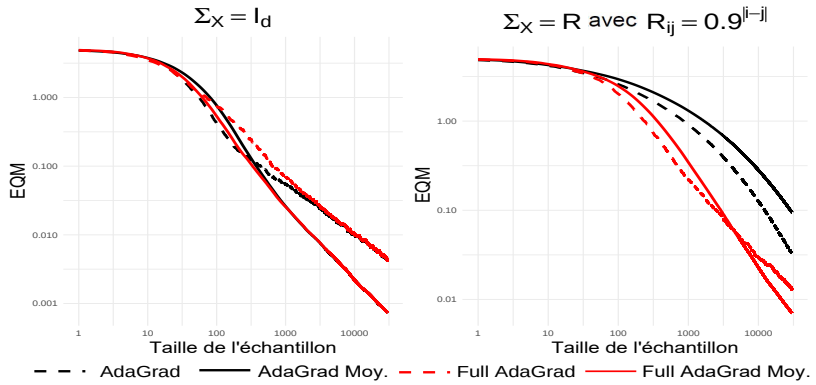


Figure 1: Cas de la régression linéaire avec $(N, d) = (30000, 20)$. Erreur quadratique moyenne en fonction de la taille de l'échantillon pour les algorithmes AdaGrad et Full AdaGrad, ainsi que leurs versions moyennées. Deux valeurs de Σ_X sont considérées : $\Sigma_X = I_d$ (à gauche) et $\Sigma_X = R$ (à droite). Les courbes en rouge représentent Full AdaGrad, tandis que celles en noir correspondent à AdaGrad. Les lignes pleines indiquent les versions moyennées des algorithmes, tandis que les lignes en pointillés représentent les versions non moyennées.

Bibliographie

Bottou, L., Curtis, F. E. et Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311.

Boyer, C. et Godichon-Baggioni, A. (2023). On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, 84(3):921–972.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K. *et al.* (2012). Large scale distributed deep networks. *Advances in neural information processing systems*, 25.

Duchi, J., Hazan, E. et Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Gadat, S. et Panloup, F. (2023). Optimal non-asymptotic analysis of the ruppert–polyak averaging stochastic algorithm. *Stochastic Processes and their Applications*, 156:312–348.

Genevay, A., Cuturi, M., Peyré, G. et Bach, F. (2016). Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29.

Godichon-Baggioni, A. et Lu, W. (2024). Online stochastic newton methods for estimating the geometric median and applications. *Journal of Multivariate Analysis*, page 105313.

Mokkadem, A. et Pelletier, M. (2011). A generalization of the averaging procedure: The use of two-time-scale algorithms. *SIAM Journal on Control and Optimization*, 49(4):1523–1543.

Pelletier, M. (1998). On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic processes and their applications*, 78(2):217–244.

Pelletier, M. (2000). Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM Journal on Control and Optimization*, 39(1):49–72.

Polyak, B. T. et Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855.

Robbins, H. et Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Seide, F., Fu, H., Droppo, J., Li, G. et Yu, D. (2014). 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. *In Fifteenth annual conference of the international speech communication association*.

Smith, L. N. (2017). Cyclical learning rates for training neural networks. *In 2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE.

Sun, S., Cao, Z., Zhu, H. et Zhao, J. (2019). A survey of optimization methods from a machine learning perspective. *IEEE transactions on cybernetics*, 50(8):3668–3681.

Zhu, W., Chen, X. et Wu, W. B. (2023). Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404.

Détection de Ruptures Multivariés en Ligne : Liens avec la Géométrie Computationnelle

Liudmila Pishchagina [1] & Gaetano Romano [2] & Paul Fearnhead [2] & Vincent Runge [1] & Guillem Rigaill [1,3]

[1] *Laboratoire de Mathématiques et Modélisation d'Evry*
liudmila.pishchagina@univ-evry.fr & vincent.runge@univ-evry.fr
[2] *Department of Mathematics and Statistics,*
Lancaster University, Lancaster LA1 4YF, UK
g.romano@lancaster.ac.uk & p.fearnhead@lancaster.ac.uk
[3] *Institute of Plant Sciences Paris-Saclay (IPS2) Orsay, France*
guillem.rigaill@inrae.fr

Résumé. L'augmentation du volume des flux de données pose des problèmes computationnels important pour la détection de ruptures en ligne. Les méthodes basées sur la vraisemblance sont efficaces, mais une mise en œuvre séquentielle naïve est impraticable en ligne en raison de son coût computationnel élevé.

Nous développons un algorithme en ligne qui calcule exactement le test du rapport de vraisemblance pour une seul rupture dans un flux de données de dimension p en tirant parti d'une connexion avec la géométrie computationnelle. Cette connexion nous permet aussi de calculer exactement des statistiques de rapport de vraisemblance sparses : c'est-à-dire en supposant que seul un sous-ensemble des p dimensions changent.

Notre algorithme est simple, rapide et apparemment quasi-linéaire. Nous démontrons qu'une variante dyadique de notre algorithme est quasi-linéaire, $\mathcal{O}(n\log(n)^{p+1})$ pour n observations et p inférieur à 3, mais plus lente en pratique. Ces algorithmes sont computationnellement impraticables lorsque p est supérieur à 5, et nous proposons une approximation adaptée à de tels p qui est $\mathcal{O}(np\log(n)^{\tilde{p}+1})$, pour un $\tilde{p}$ spécifié par l'utilisateur tel que $\tilde{p} \leq 5$. Nous dérivons certaines garanties statistiques pour les procédures proposées dans le cas gaussien, et confirmons les bonnes performances computationnelles et statistiques, ainsi que l'utilité, des algorithmes sur des simulations et sur des données de la NBA.

Mots-clés. Détection de ruptures, En ligne

Abstract. The increasing volume of data streams poses significant computational challenges for detecting changepoints online. Likelihood-based methods are effective, but a naive sequential implementation becomes impractical online due to high computational costs. We develop an online algorithm that exactly calculates the likelihood ratio test for a single changepoint in p-dimensional data streams by leveraging fascinating connections with computational geometry. This connection straightforwardly allows us to recover sparse likelihood ratio statistics exactly: that is assuming only a subset of the dimensions are changing. Our algorithm is straightforward, fast, and apparently quasi-linear. A dyadic variant of our algorithm is provably quasi-linear, being $\mathcal{O}(n\log(n)^{p+1})$ for n data points and p less than 3, but

slower in practice. These algorithms are computationally impractical when p is larger than 5, and we provide an approximate algorithm suitable for such p which is $\mathcal{O}(np\log(n)^{\tilde{p}+1})$, for some user-specified $\tilde{p} \leq 5$. We derive some statistical guarantees for the proposed procedures in the Gaussian case, and confirm the good computational and statistical performance, and usefulness, of the algorithms on both empirical data and on NBA data.

Keywords. Changepoint detection, Online

In recent years, many methods have been proposed for detecting one or multiple changepoints offline or online in data streams. The reason for such a keen interest in changepoint detection methods lies in its importance for various real-world applications, including bioinformatics (Olshen et al. (2004) and Picard et al. 2005), econometrics (Bai and Perron (2003)), medicine (Bosc et al. (2003)), climate (Reeves et al. (2007), finance (Andreou and Ghysels 2002), or neuroscience (Jewell et al. 2020). Whether this is online (i.e. observations are processed as they become available) or offline (i.e. observations are processed only once the whole data is available) the ever-increasing size of the data stream raises computational challenges. Ideally, a changepoint procedure should be linear or quasi-linear in the size of the data and, for online applications, should also be able to process new observations sequentially, and as quickly as they arrive. A common idea for both offline and online parametric (multiple) changepoint model estimation, is to optimise a likelihood function or calculate likelihood ratio statistics (Siegmund and Venkatraman (1995), Auger and Lawrence (1989)). From a computational perspective this optimisation can either be done (1) heuristically using, for example, Page CUSUM (Page (1954)) in an online context or Binary Segmentation and its variants (Scott (1974) and Fryzlewicz (2014)) in an offline context, or (2) exactly using dynamic programming techniques as in the Segment Neighbourhood algorithm (Auger and Lawrence (1989)). Importantly, these later algorithms used with an appropriate penalty enjoy good statistical properties (Yao (1984), Lavielle and Moulines (2000), Lebarbier (2005), Garreau and Arlot (2018)) and have shown good properties in numerous applications (Lai et al. (2005), Liehrmann et al. (2021)).

Pruning ideas have emerged in the literature to speed up these dynamic programming algorithms without sacrificing their exactness (Killick et al. (2012), Rigaill (2015), Maidstone et al. (2017) and Romano et al. (2022)). Two types of rules have been proposed. Inequality pruning rules, as in the PELT algorithm (Killick et al. (2012)), apply to a wide range of changepoint models but are efficient only if the number of true changepoints increases quickly with the size of the data. Functional pruning, as in the pDPA, FPOP or FOCuS (Rigaill (2015), Maidstone et al. (2017) and Romano et al. (2022)) algorithms, applies mostly to univariate models but is efficient even when the number of true changepoints is small in comparison to the size of the data. Interestingly, these algorithms are sequential in nature and empirically quasi-linear. This quasi-linearity was recently proven for the online detection of a single changepoint (Romano et al. (2022)). The pDPA, FPOP, and FOCuS algorithms rely on functionalisation of the problem, that is they consider the likelihood as a function of the last segment parameter. A direct extension of these ideas to a higher dimension ($p > 1$) has proven difficult because it involves computing the intersection and subtraction of many convex sets in $\mathbb{R}^p$ (Runge (2020), Pishchagina et al. (2024)).

In this work, we study online models with one changepoint, where the likelihood can be algebraically written as a log-likelihood from the p-variate natural exponential family. This includes, for example, the Gaussian, Poisson, categorical and Pareto type-I distribution (assuming the minimum value is known) and also non-parametric e-detectors (Shin et al. (2022)). We extend the functional pruning approach to such p-variate models using a linearisation argument embedding all likelihoods into a bigger space of dimension $p+2$. This idea allows us to demonstrate a link between the functional problem and problems studied in computational geometry.

In particular, we show that the solutions of this extension correspond to identifying the supporting hyperplanes of a particular half-space intersection problem and for models with one changepoint demonstrate that this problem further simplifies and is dual to a lower dimensional convex hull problem of dimension $p+1$. In this later case, we further prove, assuming i.i.d data, that the number of points on this hull at time n is asymptotically $\mathcal{O}(\log^p(n))$. Leveraging this leads to an empirically log-linear per iteration algorithm for calculating the likelihood ratio test statistics for a change at t.

At any time-step of this algorithm, we can exactly compute the maximum likelihood ratio-statistics and maximum sparse likelihood ratio-statistics allowing only s coordinates to change with $s \leq p$. We derive non-asymptotic bounds of these statistics to control the false alarm rate and derive a bound on the expected detection delay. While the computational cost is quasi-linear, our exact algorithms become impractical for large p, due to the $\log^p(n)$ scaling. For such cases we provide an approximation storing only an order of $\mathcal{O}(p\log^{\tilde{p}}(n))$ candidates on the hull, where $\tilde{p} \leq p$ is chosen by the user.

More details can be found in the arxiv Pishchagina et al. (2024).

Bibliographie

Adam Olshen, E.S. Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. Biostatistics (Oxford, England), 5:557–72, 11 2004.

Franck Picard, Stephane Robin, Marc Lavielle, Christian Vaisse, and Jean-Jacques Daudin. A statistical approach for array cgh data analysis. BMC Bioinformatics, 6:np, 2005.

Jushan Bai and Pierre Perron. Computation and analysis of multiple structural change models. Journal of Applied Econometrics, 18(1):1–22, 2003.

Alexander Aue, Lajos Horváth, Marie Hušková, and Piotr Kokoszka. Change-point monitoring in linear models. The Econometrics Journal, 9(3):373–403, 2006.

Marcel Bosc, Fabrice Heitz, Jean-Paul Armspach, Izzie Jacques Namer, Daniel Gounot, and Lucien Rumbach. Au- tomatic change detection in multimodal serial mri: application to multiple sclerosis lesion evolution. NeuroImage, 20:643–656, 2003.

Jaxk Reeves, Jien Chen, Xiaolan L. Wang, Robert Lund, and Qi Qi Lu. A review and comparison of changepoint detection techniques for climate data. Journal of Applied Meteorology

and Climatology, 46(6):900 – 915, dec 2007.

Elena Andreou and Eric Ghysels. Detecting multiple breaks in financial market volatility dynamics. Journal of Applied Econometrics, 17(5):579–600, 2002.

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. The Annals of Statistics, 42(6), dec 2014.

Sean Jewell, Paul Fearnhead, and Daniela Witten. Testing for a change in mean after change-point detection., 2019.

David Siegmund and ES Venkatraman. Using the generalized likelihood ratio statistic for sequential detection of a change-point. The Annals of Statistics, pages 255–271, 1995.

Ivan E. Auger and Charles E. Lawrence. Algorithms for the optimal identification of segment neighborhoods. Bulletin of Mathematical Biology, 51(1):39–54, Jan 1989

Ewan S Page. Continuous inspection schemes. Biometrika, 41(1/2):100–115, 1954.

Andrew Jhon Scott and M Knott. A cluster analysis method for grouping means in the analysis of variance. Biometrics, pages 507–512, 1974.

Yi-Ching Yao. Estimation of a noisy discrete-time step function: Bayes and empirical bayes approaches. The Annals of Statistics, 12(4):1434–1447, 12 1984.

Marc Lavielle and Eric Moulines. Least-squares estimation of an unknown number of shifts in a time series. Journal of time series analysis, 21(1):33–59, 2000.

Emilie Lebarbier. Detecting multiple change-points in the mean of gaussian process by model selection. Signal Processing, 85:717–736, 04 2005.

Damien Garreau and Sylvain Arlot. Consistent change-point detection with kernels, 2017.

Weil R Lai, Mark D Johnson, Raju Kucherlapati, and Peter J Park. Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. Bioinformatics, 21(19):3763–3770, 2005.

Arnaud Liehrmann, Guillem Rigaill, and Toby Dylan Hocking. Increased peak detection accuracy in over-dispersed chip-seq data with supervised segmentation models. BMC bioinformatics, 22(1):1–18, 2021.

Rebecca Killick, Paul Fearnhead, and Idris A. Eckley. Optimal detection of changepoints with a linear computa- tional cost. Journal of the American Statistical Association, 107(500):1590–1598, 2012.

Guillem Rigaill. A pruned dynamic programming algorithm to recover the best segmentations with 1 to K_{max} change-points. Journal de la Société Française de Statistique, 156(4):180–205, 2015.

Robert Maidstone, Toby Hocking, Guillem Rigaill, and Paul Fearnhead. On optimal multiple changepoint algorithms for large data. Statistics and Computing, 27(2):519–533, Mar 2017.

Gaetano Romano, Idris Eckley, Paul Fearnhead, and Guillem Rigaill. Fast online changepoint detection via functional pruning cusum statistics, 2022.

Vincent Runge. Is a finite intersection of balls covered by a finite union of balls in euclidean spaces? Journal of Optimization Theory and Applications, 187(2):431–447, 2020.

Liudmila Pishchagina, Guillem Rigaill, and Vincent Runge. Geometric-based pruning rules for change point detection in multiple independent time series. Computo, 2024.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. E-detectors: a nonparametric framework for sequential change detection. arXiv preprint arXiv:2203.03532, 2022.

Liudmila Pishchagina, Gaetano Romano, Paul Fearnhead, Vincent Runge and Guillem Rigaill. Online Multivariate Changepoint Detection: Leveraging Links With Computational Geometry. arXiv preprint arXiv:2311.01174 (2023).

DUST: Duality-Based Pruning Methods For Exact Multiple Change Point Detection

Vincent Runge [1] & Charles Truong [2] & Simon Querné [3]

[1] *Laboratoire de Mathématiques et Modélisation d'Évry, Université d'Évry Paris-Saclay. France. vincent.runge@univ-evry.fr*
[2] *Centre Borelli, ENS Paris-Saclay. France. charles.truong@ens-paris-saclay.fr*
[3] *Laboratoire de mathématiques de Versailles, UVSQ. France. simon.querne@ens.uvsq.fr*

Résumé. Nous abordons le problème de la détection de points de rupture multiples dans de longues séries temporelles en optimisant une vraisemblance pénalisée dérivée de modèles de la famille exponentielle. L'algorithme de programmation dynamique permet de résoudre ce problème de manière exacte, avec une complexité en temps au plus quadratique. Au cours de la dernière décennie, des méthodes d'élaguage ont été développées pour améliorer leur efficacité computationnelle. Cependant, les techniques existantes présentent des limitations notables : PELT montre une efficacité réduite en cas de ruptures peu nombreuses, tandis que la structure de FPOP est mal adaptée aux modèles multi-paramétriques. Pour surmonter ces limitations, nous introduisons ici DUST (DUalité Simple Test), une classe polyvalente de méthodes d'élaguage permettant d'améliorer l'efficacité computationnelle sur une large gamme de modèles. Le test d'élagage de DUST repose sur l'évaluation d'une fonction duale comparée à un seuil, ce qui le rend applicable à tout modèle paramétrique, quelle que soit sa dimension. Nous montrons l'efficacité de DUST sur différents régimes de ruptures et types de modèles. Enfin, nous appliquons DUST à des données EEG dans un modèle à changement de variance, illustrant sa capacité à inférer efficacement la structure optimale des ruptures.

Mots-clés. Séries temporelles, Détection de ruptures, Programmation dynamique, Complexité algorithmique

Abstract. We tackle the challenge of detecting multiple change points in large time series by optimizing a penalized likelihood derived from exponential family models. Dynamic programming algorithms can solve this problem exactly, with at most quadratic time complexity. Over the past decade, pruning methods have been developed to improve their efficiency. However, existing techniques have notable limitations: PELT struggles with pruning efficiency in sparse-change scenarios, while FPOP's structure is ill-suited for multiparametric settings. To address these issues, we introduce DUST (DUality Simple Test), a versatile class of pruning methods that enhances computational efficiency across various models. DUST's pruning test is based on evaluating a dual function against a threshold, making it broadly applicable to parametric models of any dimension. We demonstrate DUST's effectiveness across different change point regimes and models. Finally, we apply DUST to EEG data using a change-in-variance model, illustrating its ability to efficiently infer the optimal change structure.

Keywords. Time series, Change point detection, Dynamic programming, Computational complexity

1 Model and optimization problem

We consider a d-dimensional multivariate time series made of n data points, $y_1, y_2, \ldots, y_n$ drawn independently from a distribution in the exponential family. The subset of points, $y_{a..b}$, is called a segment and is defined as the collection of consecutive data points $(y_{a+1}, \ldots, y_b)$. One can define its associated cost function over such a segment using its sufficient statistic function $\mathbf{S}_{a..b} = \sum_{i=a+1}^{b} \mathbf{T}(y_i) \in \mathbb{R}^d$ as:

$$c(y_{a..b}; \boldsymbol{\theta}) = (b-a)A(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot \mathbf{S}_{a..b}\,,$$

where the convex function A is called the log-partition function and parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T$ is a d-dimensional point.

Example 1 *For univariate data we write $S_{a..b} = \sum_{i=a+1}^{b} y_i$. With a Poisson model, its one-parametric cost function is given by $c(y_{a..b}; \theta) = (b-a)\exp(\theta) - S_{a..b}\theta$ while for a Gaussian distribution with unitary variance, $c(y_{a..b}; \theta) = (b-a)\frac{\theta^2}{2} - S_{a..b}\theta$.*

We perform multiple change point detection via penalized maximum likelihood, where the penalization term is linear in the number of segments. To each segment, we can associate a positive value β called the (unitary) penalty. The greater this value, the sparser is the obtained segmentation. We define the optimum penalized log-likelihood as follows:

$$Q_n = \min_{\tau \in \mathcal{T}} \left(\sum_{k=0}^{K} \Big[\min_{\boldsymbol{\theta}} c(y_{\tau_k..\tau_{k+1}}, \boldsymbol{\theta}) + \beta \Big] \right)\,,$$

where $\tau = (\tau_0 = 0, \tau_1, \ldots, \tau_K, \tau_{K+1} = n)$ is a change point vector belonging to the set of all available segmentations $\mathcal{T} = \{\tau \in \mathbb{N}^{K+2}, K \in \mathbb{N}, 0 = \tau_0 < \tau_1 < \ldots < \tau_K < \tau_{K+1} = n\}$. Note that K is also an unknown parameter, whose value depends on the choice of penalty value.

We compute the optimal solution using a dynamic programming approach, finding all the Q_t values for $t = 1, \ldots, n$. To that end, we search for the best last segment using the optimal partitioning algorithm [1]:

$$Q_t = \min_{s \in S_t} \Big\{ Q_s + \min_{\boldsymbol{\theta}} c(y_{s..t}; \boldsymbol{\theta}) + \beta \Big\}\,,$$

where S_t is a set of indices, a subset of $\{0, 1, \ldots, t-1\}$. If no pruning rule is applied, then $S_t = \{0, 1, \ldots, t-1\}$ and the time complexity is quadratic. The game we play in dynamic programming consists of finding pruning rules allowing a reduction of the number of elements in S_t, while still keeping the solution of the dynamic programming exact. This is made possible using a simple inequality-based rule called PELT [2] or a functional-based method called FPOP [3].

In our work, we propose some new duality-based pruning rules called DUST allowing us to drastically reduce the number of indices to consider in S_t. This rule is a natural generalization of the PELT rule. Unlike PELT, this rule is also efficient in the no-change regime; unlike FPOP, we can easily use the DUST rule in multivariate settings.

2 Functional pruning seen as an optimization problem

The optimal partitioning rule can be written as:

$$Q_t(\boldsymbol{\theta}) = \min_{s \in S_t} \left\{ Q_s + c(y_{s..t}; \boldsymbol{\theta}) + \beta \right\} = \min_{s \in S_t} \left\{ q_t^s(\boldsymbol{\theta}) \right\},$$

where $Q_t = \min_{\boldsymbol{\theta}} Q_t(\boldsymbol{\theta})$. We have simply exchanged the order of the two min operators, deferring the minimization over the continuous parameter $\boldsymbol{\theta}$ to the final step.

In the evolution of the dynamic programming algorithm, the inner functions $q_t^s(\cdot)$ are discarded only by the last introduced function at time $t+1$, that is $Q_t + c(y_{t..(t+1)}; \boldsymbol{\theta}) + \beta$. The natural strategy is then the following: comparing functions at time t with the constant $Q_t + \beta$. We denote by Θ_t^s the feasible set of visible values for inner cost $q_t^s(\cdot)$ in functional cost function $Q_t(\cdot)$:

$$\Theta_t^s := \{\boldsymbol{\theta} | \forall r \neq s,\ q_t^s(\boldsymbol{\theta}) \leq q_t^r(\boldsymbol{\theta})\}. \tag{1}$$

A method based on approximating the set Θ_t^s by simple shapes has been recently proposed [4]. However, its approximation is difficult for non-Gaussian cost and its update is relatively slow (see also [5]). Instead, we propose to look at the minimal value of q_t^s in this set built as the sequential comparison with levels "$Q_u + \beta$".

Proposition 1 (Functional pruning rule) *We consider Θ_t^s in (1) and $R_t^s := \min_{\boldsymbol{\theta} \in \Theta_t^s} q_t^s(\boldsymbol{\theta})$. By convention, if $\Theta_t^s = \emptyset$, then $R_t^s = +\infty$. The functional pruning condition is: if there exists s in S_t satisfying:*

$$R_t^s > Q_t + \beta, \tag{2}$$

then s can never be the last change point of a segmentation of $y_{0..t'}$: for all $t' > t$ that is $s \notin S_{t'}$.

The quantity R_t^s in Proposition 1 is not easy to compute. We can explicitly see the general problem of pruning for multiple change point detection as a specific problem of optimization under constraints.

Definition 1 (Pruning problem) *The pruning problem for multiple change point detection at time t for testing change point s is the following problem of optimization under constraints:*

$$PR(t) : \begin{cases} \min\limits_{\boldsymbol{\theta}} \left(q_t^s(\boldsymbol{\theta}) \right) \\ s.t.\ q_t^s(\boldsymbol{\theta}) - q_t^r(\boldsymbol{\theta}) \leq 0 \quad \textit{for all} \quad r \neq s \in \{0, \ldots, t-1\}. \end{cases} \tag{3}$$

A minimum bigger than the threshold $Q_t + \beta$ is equivalent to pruning.

Under some mild conditions (choosing only some of the available constraints), this optimization problem (3) can be solved in Lagrangian duality with no duality gap. The obtained dual function $\mathcal{D}$ is given under closed form; this is a concave function and its maximum is

exactly the solution of (3). Note that any evaluation of the dual is a lower bound of (3), thus avoiding any spurious pruning.

We propose here a simple DUST rule using only one constraint.

Proposition 2 (DUST pruning rule) *At time t we prune index s using candidate index $r < s$ if $\mathcal{D}(\mu_0) > Q_t + \beta$ where $\mu_0 \sim \mathcal{U}[0, \mu_{max}]$ and $\mathcal{D}$ is the dual function of the optimization problem* (3).

3 Example of the change in mean and variance problem

We present the DUST method on a specific change point problem. We choose the well-known change in mean and variance problem for Gaussian models. To the best of our knowledge, there are no available efficient implementations of this problem as it faces two difficulties. First, we deal with a two-parametric problem (FPOP cannot be used easily [3]). Second, the logarithmic term in the cost function leads to difficult computations for finding roots (geomFPOP cannot be used easily [4]). With one constraint $r < s$ only, we obtain the following dual function at time t:

$$\mathcal{D}(\mu) = \frac{1}{2}\Big((t-s) - \mu(s-r)\Big)\left[1 + \log\Big(\frac{S^2{}_{s..t} - \mu S^2{}_{r..s}}{(t-s)-\mu(s-r)} - \Big(\frac{S_{s..t} - \mu S_{r..s}}{(t-s)-\mu(s-r)}\Big)^2\Big)\right]$$
$$+\beta + (1+\mu)Q_s - \mu Q_r\,,$$

where $S_{r..s} = \sum_{i=r+1}^{s} y_i$ and $S^2_{r..s} = \sum_{i=r+1}^{s} y_i^2$.

In Figure 1 we evaluate the efficiency of DUST for this example. For each index s to be considered for the potential last change point, we randomly sample one index r (smaller than the one considered) and pick a random uniform position on $[0, \mu_{max}]$ for the dual. With $n = 10^4$, $\beta = 4\log(n)$ and data with no change, only 4.4% of the indices remain unpruned and the overall number of indices to be considered is reduced by a factor 20 compared to PELT (left panel). With changes located at one third and two thirds of the data, with segment model $(\mathcal{N}(0,1), \mathcal{N}(0,4), \mathcal{N}(0.5,4))$ DUST is still also efficient in terms of pruning capacity (right panel). Running times with dust R-code package are clear, for $n = 10^5$ and no change, $T(DUST) = 4.7s$ while $T(PELT) = 82s$: it is important to highlight the fact that the pruning cost of the DUST test is of the same order as the PELT test.

An Rcpp package for DUST can be found here `https://github.com/vrunge/dust`. It proposes DUST analysis for many multivariate model of the exponential family (7 different models). A pre-print on arXiv will be available soon. The method is already presented in a specific parametric case in [6]. Many other simulation results will be presented during the talk.

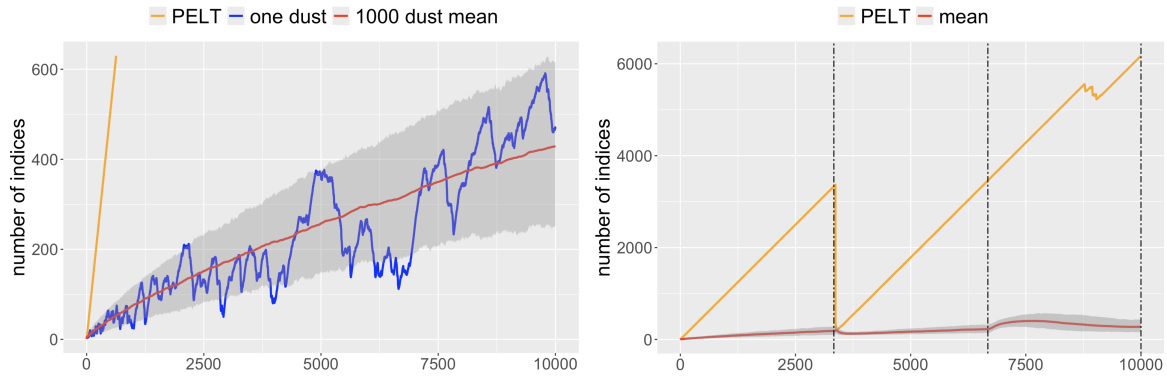


Figure 1: (left) For data with no change the number of indices saved by dust over time is consistently much less than PELT, even in the presence of changes (right)

Bibliographie

[1] Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. (2005). *An algorithm for optimal partitioning of data on an interval. IEEE Signal Processing Letters*, **12**(2), 105–108.

[2] Killick, R., Fearnhead, P., and Eckley, I. A. (2012). *Optimal detection of changepoints with a linear computational cost. Journal of the American Statistical Association*, **107**(500), 1590–1598.

[3] Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2017). *On optimal multiple changepoint algorithms for large data. Statistics and Computing*, **27**, 519–533.

[4] Pishchagina, L., Rigaill, G., and Runge, V. (2024). *Geometric-based pruning rules for change point detection in multiple independent time series. Computo.* Retrieved from `https://openreview.net/forum?id=YIes47zsCE`

[5] Runge, V. (2020). *Is a finite intersection of balls covered by a finite union of balls in Euclidean spaces? Journal of Optimization Theory and Applications*, **187**(2), 431–447.

[6] Truong, C., and Runge, V. (2024). *An efficient algorithm for exact segmentation of large compositional and categorical time series. Stat*, **13**(4), e70012.

Statistique des processus

Estimation adaptative du coefficient de diffusion dans les EDS

Yichuan Huang[1,2]

[1] *Laboratoire Modal'X, Université Paris Nanterre, France,* yhuang@parisnanterre.fr
[2] *Laboratoire MAP5, Université Paris Cité, France*

Résumé. Nous étudions un estimateur non paramétrique du coefficient de diffusion σ^2 dans une équation différentielle stochastique (EDS) à partir d'observations en haute fréquence et à long terme. Notre approche se fonde sur des estimations par projection orthogonale dans des espaces de dimension finie, complétées par une procédure de sélection de modèle pénalisée pour choisir automatiquement la dimension optimale de projection. Nous présentons ensuite des bornes de risque, valables même pour des domaines non compacts (par exemple, des bases de Laguerre ou d'Hermite). Enfin, nous montrons, sur plusieurs processus de diffusion (Ornstein-Uhlenbeck, CIR, etc.), que cette méthode offre un bon compromis entre le biais et la variance et demeure numériquement stable pour de grands échantillons.

Mots-clés. Processus de diffusion, Estimation non paramétrique, Sélection de modèle, Projection orthogonale, Mélange exponentiel.

Abstract. We study a nonparametric estimator for the diffusion coefficient σ^2 in a stochastic differential equation (SDE) based on high-frequency and long-term observations. Our approach is based on orthogonal projection estimators in finite-dimensional spaces, complemented by a penalized model selection procedure to automatically choose the optimal projection dimension. We then present risk bounds that remain valid even for noncompact domains (e.g., Laguerre or Hermite bases). Finally, we demonstrate on several diffusion processes (Ornstein-Uhlenbeck, CIR, etc.) that this method offers a good trade-off between bias and variance and remains numerically stable for large sample sizes.

Keywords. Diffusion processes; Nonparametric estimation; Model selection; Orthogonal projection; Exponential mixing.

1 Introduction

Les équations différentielles stochastiques (EDS) jouent un rôle majeur pour modéliser des phénomènes aléatoires en finance (taux d'intérêt, volatilité), en biologie (croissances cellulaires, dynamiques de populations) ou en physique (processus de diffusion, etc.). Nous considérons le modèle

$$dX_t \;=\; b(X_t)\,dt \;+\; \sigma(X_t)\,dW_t, \quad t \geq 0, \tag{1}$$

où $(W_t)_{t\geq 0}$ est un mouvement brownien standard. Le drift b et la diffusion σ sont inconnus. Notre but est d'estimer $\sigma^2(\cdot)$ à partir d'observations discrètes $\{X_{k\Delta}\}_{k=0,\dots,n}$.

Méthodes existantes. Classiquement, deux grandes familles de méthodes ont été proposées :

1. *Estimateurs à noyau.* Un estimateur proposé par Florens-Zmirou (1993), qui utilise une fenêtre δ_n et permet de construire une estimation ponctuelle. Par la suite, plusieurs travaux ont généralisé ou raffiné cette approche, notamment Jacod (2000), Jiang & Knight (1997) et Bandi & Phillips (2003).

2. *Estimateurs par projection.* Plusieurs contributions (Hoffmann (1999), Comte *et al.* (2007), Comte & Genon-Catalot (2021)) traitent de l'approximation de la fonction $\sigma^2(\cdot)$ dans un espace de dimension m, avant de sélectionner m à partir des données. Mentionnons également des travaux plus récents pour des domaines non bornés, tels qu'Ella-Mintsa (2024).

Dans le contexte d'une EDS ergodique, la consistance ainsi que des vitesses de convergence satisfaisantes sont atteignables. Notre contribution vise à **(i)** généraliser l'estimateur des moindres carrés par projection à des domaines potentiellement non compacts (bases de Laguerre ou d'Hermite), **(ii)** construire une procédure adaptative de sélection de la dimension m via une pénalité, et **(iii)** présenter des résultats numériques encourageants.

2 Cadre, hypothèses et notations

2.1 Conditions : stationnarité et mélange

Nous supposons que (1) admet une unique solution forte stationnaire $(X_t)_{t\geq 0}$, c'est-à-dire que X_0 est distribué selon la mesure invariante. De plus, cette solution est ergodique et satisfait une condition de β-mélange exponentiel : $\beta_X(t) \leq Ke^{-\theta t}$. Typiquement, on impose les conditions suivantes :

- Les fonctions b et σ sont Lipschitz, et σ est non dégénérée
- Une condition dissipative sur b (qui exerce une force de rappel vers la mesure invariante)
- Un échantillonnage haute fréquence, c'est-à-dire $\Delta_n \to 0$ et $n\Delta_n \to \infty$.

2.2 Notations

On note

$$U_k = \frac{(X_{(k+1)\Delta} - X_{k\Delta})^2}{\Delta}, \quad k = 0, \ldots, n-1.$$

En outre, pour des fonctions f et g, on définit le produit scalaire empirique et la norme associée par

$$\langle f, g\rangle_n = \frac{1}{n}\sum_{k=1}^{n} f\big(X_{k\Delta}\big)\, g\big(X_{k\Delta}\big), \quad \|f\|_n^2 = \langle f, f\rangle_n,$$

et la norme $L^2(\pi)$ via $\|f\|_\pi^2 = \mathbb{E}[f(X_0)^2]$ (où $X_0 \sim \pi(x)dx$).

3 Estimateur des moindres carrés par projection

3.1 Étape de projection

Soit $(\varphi_j)_{j\geq 0}$ une famille orthonormale dans $L^2(A)$ (où A est le support, potentiellement non compact). Pour un entier $m \geq 1$, on considère

$$S_m = \operatorname{span}\{\varphi_0, \ldots, \varphi_{m-1}\}.$$

On définit alors

$$\widehat{\sigma}_m^2 = \arg\min_{t \in S_m} \frac{1}{n} \sum_{k=1}^{n} \Big(U_k - t\big(X_{k\Delta}\big)\Big)^2.$$

Un estimateur du coefficient inconnu σ^2 est ainsi obtenu par projection sur un espace de dimension m. Suivant Comte et Genon-Catalot (2021), on applique une troncature :

$$\widetilde{\sigma}_m^2 = \widehat{\sigma}_m^2 \, \mathbf{1}_{\Lambda_m},$$

où

$$\Lambda_m := \Big\{ L(m)\big(\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}} \vee 1\big) \leq \mathfrak{c} \frac{n\Delta}{\log^2(n\Delta)} \Big\}, \quad \widehat{\Psi}_m = \Big[\langle \varphi_j, \varphi_k \rangle_n\Big]_{0\leq j,k\leq m-1},$$

$$\mathfrak{c} = \frac{\theta\big(3\log(3/2) - 1\big)}{82}.$$

3.2 Sélection de m par critère pénalisé

Pour choisir m automatiquement, on introduit d'abord une collection de modèles théoriques $\mathcal{M}_{n\Delta} \subset \mathbb{N}$ définie par :

$$\mathcal{M}_{n\Delta} := \left\{ m \in \mathbb{N}^+ : c_\varphi^2 \, m \, (\|\Psi_m^{-1}\|_{\mathrm{op}}^2 \vee 1) \leq \frac{\mathfrak{d}}{4} \frac{n\Delta}{\log^2(n\Delta)} \right\}, \quad \mathfrak{d} = \frac{\theta}{672(\|\pi\|_\infty \vee 4/3)}.$$

Cette contrainte, analogue à celle proposée par Cohen et al. (2013), permet d'assurer la stabilité numérique liée à l'inversion de la matrice $\Psi_m = \mathbb{E}[\widehat{\Psi}_m]$, matrice de Gram théorique.

Ensuite, on considère une version empirique $\widehat{\mathcal{M}}_{n\Delta}$ de cette collection et on pose :

$$\widehat{m} = \arg\min_{m \in \widehat{\mathcal{M}}_{n\Delta}} \left(-\|\widehat{\sigma}_m^2\|_n^2 + \widehat{\mathrm{pen}}(m) \right),$$

où la pénalité est définie précisément par :

$$\widehat{\mathrm{pen}}(m) = \kappa \, c_\varphi^2 \, s^2 \, \frac{m \|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}}{n}, \quad s^4 = \mathbb{E}[\sigma^4(X_0)],$$

où c_φ est une constante satisfaisant

$$L(m) \le c_\varphi^2\, m, \quad \text{avec} \quad L(m) = \sup_{x\in A} \sum_{j=0}^{m-1} \varphi_j(x)^2,$$

et κ est une constante numérique théoriquement grande, mais pouvant être calibrée numériquement.

3.3 Borne de risque

Nous pouvons énoncer précisément le résultat principal :

Proposition 1 (Borne de risque adaptative)**.** *Sous les hypothèses de stationnarité et de β-mélange exponentiel de la Section 2, supposons que les espaces de projection satisfont $S_{m_1} \subseteq S_{m_2}$ pour $m_1 \le m_2$, et $L(m) \le c_\varphi^2 m$. Alors, l'estimateur adaptatif $\widehat{\sigma}^2_{\widehat{m}}$ satisfait :*

$$\mathbb{E}[\|\widehat{\sigma}^2_{\widehat{m}} - \sigma_A^2\|_n^2] \le C \inf_{m\in\mathcal{M}_{n\Delta}} \left(\inf_{t\in S_m} \|t - \sigma_A^2\|_\pi^2 + \mathrm{pen}(m) \right) + C'\Delta^2 + C''\frac{1}{n},$$

où $C, C', C'' > 0$ sont des constantes indépendantes de n, m, dès que $\Delta_n \to 0$, $\Delta_n n^{3/5} \to \infty$ et $\Delta_n^2 = o(1/n)$ lorsque $n \to \infty$.

Remarque. L'obtention de cette borne nécessite des outils probabilistes sophistiqués (inégalités de Talagrand), notamment pour contrôler uniformément la norme opérateur $\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}$. Cette difficulté fondamentale justifie l'importance de la procédure adaptative proposée, assurant stabilité numérique et efficacité théorique.

4 Expériences numériques

Pour évaluer les performances de l'estimateur, nous avons procédé à diverses simulations, par exemple :

4.1 Processus CIR

$$dV_t = \kappa\,(\alpha - V_t)\,dt + \omega\,\sqrt{(V_t)_+}\,dW_t,$$

ce processus est classique en finance (taux courts). Sous la condition $2\kappa\alpha \ge \omega^2$, la variable V_t reste strictement positive. Nous utilisons une base de Laguerre et la troncature correspondante pour éviter les inversions de matrices mal conditionnées, afin de construire un estimateur sur $\mathbb{R}_+$. Nous avons répété l'expérience en faisant varier n et en diminuant Δ. Les MISE obtenues restent de l'ordre de 0.001 pour de grands n. La dimension m est choisie automatiquement autour de 4–5 afin de limiter la variance.

4.2 Diffusion hyperbolique

$$dU_t = -\lambda\,(U_t - \mu)\,dt + \gamma\,\sqrt{1 + U_t^2}\,dW_t.$$

Le coefficient $\sigma(u) = \gamma\,\sqrt{1+u^2}$ augmente lorsque $|u| \to \infty$. On utilise ici une base d'Hermite sur $\mathbb{R}$ et la troncature correspondante pour éviter les inversions de matrices mal conditionnées. Les résultats obtenus sont également très satisfaisants. La dimension m est choisie automatiquement autour de 5–6 afin de limiter la variance.

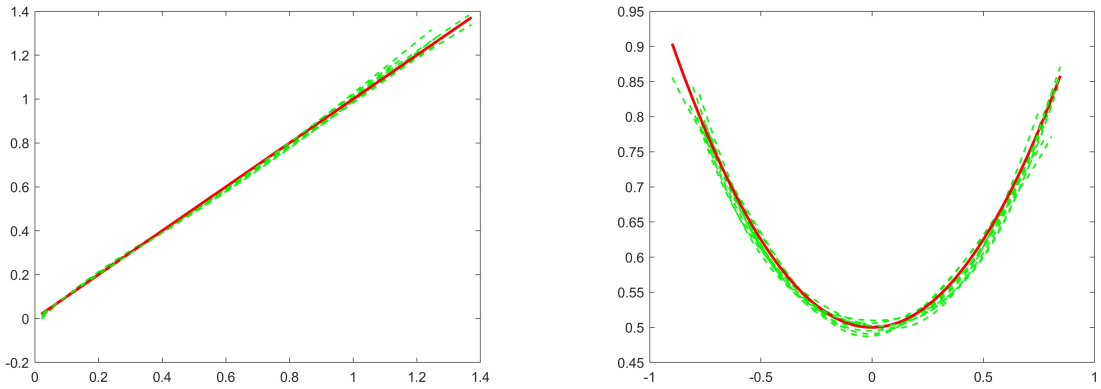


Figure 1: Exemples de reconstructions pour le processus CIR (à gauche) et la diffusion hyperbolique (à droite).

Bibliographie

Bandi, F.M. & Phillips, P.C.B. (2003). Fully nonparametric estimation of scalar diffusion models, *Econometrica*, 71(1), 241–283.

Cohen, A., Davenport, M. A. & Leviatan, D. (2013). On the stability and accuracy of least squares approximations, *Foundations of Computational Mathematics*, 13, 819–834.

Comte, F., Genon-Catalot, V. et Rozenholc, Y. (2007). Penalized projection estimators of the drift for diffusion processes, *Bernoulli*, 13(2), 514–543.

Comte, F. et Genon-Catalot, V. (2021). Nonparametric estimation for continuous-time processes: an overview, *TEST*, 30, 22–67.

Ella-Mintsa, E. (2024). Nonparametric estimation of the diffusion coefficient from iid SDE paths, *Statistical Inference for Stochastic Processes*, 1–56.

Florens-Zmirou, D. (1993). On estimating the diffusion coefficient from discrete observations, *J. Appl. Probab.*, 30, 790–804.

Hoffmann, M. (1999). Adaptive estimation in diffusion processes, *Stochastic Process. Appl.*, 79(1), 135–163.

Huang, Y. (2025). Model Selection for Diffusion Coefficient Estimation in SDE. Submitted, *hal-04901917*.

Jacod, J. (2000). Non-parametric Kernel Estimation of the Coefficient of a Diffusion, *Scand. J. Statist.*, 27(1), 83–96.

Jiang, G.J. & Knight, J.L. (1997). A nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest rate model, *Econometric Theory*, 13(5), 615–645.

Algorithme de Classification ERM-Lasso pour les Processus de Hawkes Multivariés

Christophe Denis [1] & Charlotte Dion-Blanc [2] & Romain E. Lacoste [3] & Laure Sansonnet [2]

[1] *SAMM, Université Paris 1 Panthéon-Sorbonne, France*
[2] *LPSM, Sorbonne Université, France*
[3] *LAMA, Université Gustave Eiffel, France, romain.lacoste@polytechnique.edu*

Résumé. On s'intéresse au problème de classification de trajectoires de processus de Hawkes multivariés (PHM) issues de plusieurs classes. Les PHM forment une famille polyvalente de processus ponctuels qui modélisent les interactions entre des individus connectés au sein d'un réseau. Les données d'entrainement observées consistent en des répétitions indépendantes de trajectoires étiquetées, lesquelles sont observées sur un intervalle de temps fixé. En outre, nous considérons le cadre de la grande dimension, ce qui signifie que la dimension du réseau peut être importante par rapport au nombre d'observations. Nous avons donc besoin de faire une hypothèse de parcimonie sur la matrice d'adjacence. Dans ce contexte, nous proposons une nouvelle méthodologie avec une étape initiale de sélection des interactions, par classe, suivie d'une étape de réajustement basée sur un critère de classification approprié. Pour récupérer le support de la matrice d'adjacence, un estimateur de type Lasso est proposé, pour lequel nous établissons des taux de convergence. Ensuite, en tirant parti du support estimé, nous construisons une procédure de classification basée sur la minimisation d'un risque L_2. Les vitesses de convergence de notre procédure de classification sont notamment fournies. Une phase de test approfondie utilisant des données synthétiques soutient les deux résultats théoriques.

Mots-clés. Processus de Hawkes, Classification, Minimisation du Risque Empirique, Grande Dimension, Lasso.

Abstract. We are interested in the problem of classifying Multivariate Hawkes Processes (MHP) paths coming from several classes. MHP form a versatile family of point processes that models interactions between connected individuals within a network. The observed learning data consist of labeled repeated and independent paths on a fixed time interval. Besides, we consider the high-dimensional setting, meaning the dimension of the network may be large *w.r.t.* the number of observations. We consequently require a sparsity assumption on the adjacency matrix. In this context, we propose a novel methodology with an initial interaction recovery step, by class, followed by a refitting step based on a suitable classification criterion. To recover the support of the adjacency matrix, a Lasso-type estimator is proposed, for which we establish rates of convergence. Then, leveraging the estimated support, we build a classification procedure based on the minimization of a L_2-risk. Notably, rates of convergence of our classification procedure are provided. An in-depth testing phase using synthetic data supports both theoretical results.

Keywords. Hawkes process, Classification, Empirical Risk Minimization, High Dimension, Lasso.

1 Introduction

Les processus de Hawkes, introduits par Hawkes (1971), forment une classe de processus ponctuel qui permettent de modéliser des séries temporelles où l'occurrence d'un évènement augmente la probabilité d'en observer un nouveau dans un futur proche. En vertu de leur nature à capturer des dynamiques auto-excitantes, les processus de Hawkes sont employés dans de nombreuses situations. En effet, bien qu'historiquement utilisés pour la modélisation de secousses sismiques Hawkes (1973), leur champ d'application s'est étendu à de nombreux autres domaines, comme les neurosciences Reynaud-Bouret et al. (2013) par exemple. La version multidimensionnelle, appelée processus de Hawkes multivariés (PHM), constitue une généralisation naturelle qui enrichit les possibilités de modélisation. En effet, en plus de modéliser les interactions auto-excitantes, un tel modèle prend en compte les interactions positives entre les individus connectés au sein d'un réseau.

Cette étude s'inscrit dans le cadre de la classification supervisée : nous supposons que l'échantillon d'apprentissage consiste en des données étiquetées *i.i.d.* où les variables explicatives sont des temps d'événements de PHM observés en temps court. Chaque classe est donc caractérisée par sa propre dynamique d'occurrence sous-jacente, et l'objectif est de les différencier. Le travail de Denis et al. (2022), introduit une procédure de classification pour des observations provenant d'un processus de Hawkes univarié où les classes sont discriminées par le noyau du processus. En particulier, cette méthode est appliquée dans Denis et al. (2024), pour modéliser les cris d'écholocation de chauves-souris enregistrés sur plusieurs sites en France.

Dans ce travail, nous généralisons l'approche développée dans Denis et al. (2022) au cadre des PHM. En outre, on se place dans le cadre de la grande dimension, signifiant ici que la dimension du réseau peut être importante par rapport au nombre d'observations. Pour composer avec la grande dimension, on impose des hypothèses de parcimonie sur la matrice d'interaction. Dans ce contexte, nous proposons une nouvelle méthodologie qui se décompose en deux temps. D'abord, une étape initiale de sélection des interactions effectives est réalisée, par classe ; laquelle est suivie d'une étape de réajustement basée sur un critère de classification approprié. Pour récupérer le support de la matrice d'adjacence, un estimateur de type Lasso est proposé, pour lequel nous établissons des vitesses de convergence. Puis, nous construisons une procédure de classification basée sur la minimisation d'un risque L_2, en tirant parti du support préalablement estimé. En particulier, on établit la vitesse de convergence de notre procédure. Enfin, pour valider la procédure, on illustre ses performances sur des données simulés à l'aide de la librairie **Sparklen** (Lacoste (2025)).

2 Estimateur Lasso pour estimation du support

On s'intéresse au processus de Hawkes linéaire exponentiel d-multivarié (PHM) observé en temps court sur $[0, T]$ où $d > 1$ est la dimension du réseau. Un tel processus est noté $N = (N_1, \ldots, N_d)$ où chaque coordonnée N_j est un processus ponctuel sur $\mathbb{R}_+^*$. Le processus de comptage N_j est le processus $(N_j(t))_{t\in[0,T]}$ tel que $N_j(t) = \sum_{\ell\geq 1} \mathbb{1}_{\{t_{j,\ell}\leq t\}}$, où

$\mathcal{T}_T := \left\{\{t_{j,\ell}\}_{\ell\in[N_j(T)]}, j \in [d]\right\}$ sont les temps de sauts du processus N sur $[0,T]$. Chaque N_j est caractérisé par son intensité conditionnelle, fonction définie pour tout $t \geq 0$ par :

$$\lambda_j(t) := \mu_j + \sum_{j'=1}^{d} \alpha_{j,j'} \int_0^t \beta e^{-\beta(t-s)} \, \mathrm{d}N_{j'}(s) = \mu_j + \sum_{j'=1}^{d} \alpha_{j,j'} \sum_{\ell : t_{j',\ell} < t} \beta e^{-\beta(t-t_{j',\ell})},$$

- $\mu = (\mu_j)_{j\in[d]}$: vecteur d'intensité exogène, modélise l'arrivée d'évènements spontanés ;
- $\alpha = (\alpha_{j,j'})_{j,j'\in[d]}$: matrice d'interaction, modélise l'influence positive de la composante j' sur la composante j ;
- $\beta \geq 0$: taux de décroissance, dicte la vitesse avec laquelle ces influences disparaissent.

On se place dans le cadre paramétrique et on considère que chaque λ_{j,θ^*} dépend d'un paramètre $\theta^* = (\mu^*, \alpha^*)$ inconnu qui appartient à la famille de paramètres suivante :

$$\Theta := \left\{\mu \in (\mathbb{R}_+^*)^d, \ \alpha \in \mathbb{R}_+^{d\times d}, \ \rho(\alpha) < 1\right\} \in \mathbb{R}_+^{d\times d+1}$$

avec $\rho(\alpha)$ le rayon spectral de α. En revanche on considère β comme un hyperparamètre et commun à toutes les composantes de N. Notre objectif est de reconstruire le support de θ^*, qu'on note $\mathrm{supp}(\theta^*)$.

Pour ce faire, on suppose qu'on dispose d'un n-échantillon *i.i.d.* $\mathcal{D}_n := \{\mathcal{T}_T^{(1)}, \ldots, \mathcal{T}_T^{(n)}\}$. De plus, on se place dans le cadre de la grande dimension : d peut être très grand par rapport à n. Afin de composer avec la grande dimension, on impose des hypothèses de parcimonie sur la matrice α^*. On considère le contraste des moindres carrés moyenné sur les répétitions :

$$R_{T,n}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{T} \sum_{j=1}^{d} \int_0^T \lambda_{j,\theta}^{(i)2}(t) \, \mathrm{d}t - \frac{2}{T} \sum_{j=1}^{d} \sum_{\ell : t_{j,\ell}^{(i)} < T} \lambda_{j,\theta}^{(i)} \left(t_{j,\ell}^{(i)} \right) \right)$$

où $\lambda_{j,\theta}^{(i)}(t)$ est définie à partir de la i-ème répétition.

Pour estimer $\mathrm{supp}(\theta^*)$, étant donné $\mathcal{D}_n$, on considère l'estimateur Lasso :

$$\hat{\theta} \in \underset{\theta\in\mathbb{R}^{d\times d+1}}{\arg\min} \left\{ R_{T,n}(\theta) + \kappa \sum_{j=1}^{d} \sum_{j'=1}^{d} |\theta_{j,j'}| \right\} \tag{1}$$

où κ est la constante de régularisation à calibrer.

Nous pouvons énoncer le résultat suivant, où l'on établit la consistance du support et la convergence de l'estimateur Lasso associé.

Théorème 1. *Soit $\kappa = n^{-1/2}\log^4(nd^2)$. On se place sous des hypothèses classiques d'incohérence mutuelle, de valeur propre minimale et de condition de signal minimal (voir Dion-Blanc et al. (2024)). Il existe un évènement Ω_n avec $\mathbb{P}(\Omega_n) \geq 1 - \dfrac{C}{n}$, sur lequel le contraste des moindres carrés pénalisé défini par l'équation 1 admet une unique solution $\hat{\theta}$ qui satisfait les propriétés suivantes*

1. $\text{supp}(\hat{\theta}) = \text{supp}(\theta^*)$ *;*

2.

$$\left\|\hat{\theta} - \theta^*\right\|_\infty \leq \frac{\Lambda_0 \max_{j=1,\ldots,d} \sqrt{|S^*_{\theta_j}|} \log^4(nd^2)}{\sqrt{n}}.$$

La preuve de ce résultat suit la méthode de construction *primal-dual witness* décrite dans le Chapitre 11 de Hastie et al. (2015).

3 Algorithme de classification ERM-Lasso

On se place dans le cadre de la classification multiclasse K avec $K \geq 2$ le nombre de classes. Soit $(\mathcal{T}_T, Y)$ le couple de variables formé par la variable explicative et son étiquette. Ici, $\mathcal{T}_T$ consiste en des temps d'évènement d'un PHM N observé sur $[0, T]$ dont la dynamique d'occurrence est dicté par son étiquette $Y \in [K]$. Plus précisément, conditionnellement à Y, N est un PHM de fonction d'intensité $\lambda_{\theta^*_Y}$. Pour chaque $j \in [d]$, la fonction d'intensité du processus N_j dépend de Y et est donné par :

$$\lambda_{\theta^*_Y, j}(t) = \mu^*_{Y,j} + \sum_{j'=1}^{d} \alpha^*_{Y,j,j'} \int_0^t \beta e^{-\beta(t-s)} \, \mathrm{d}N_{j'}(s)$$

Les classes sont donc discriminées par les vecteurs $(\mu^*_Y)_{Y\in[K]}$ ainsi que par les matrices $(\alpha^*_Y)_{Y\in[K]}$. Dans ce qui suit, on suppose que les paramètres $(\mu^*_k, \alpha^*_k)_{k\in[K]}$ ainsi que la loi de Y qui est notée $(p^*_k)_{k\in[K]}$ sont inconnus, tandis que β est supposé fixé.

Nous suivons le paradigme de la classification supervisée : alors que la loi de la paire $(\mathcal{T}_T, Y)$ est inconnue, on suppose que l'on a accès à un n-échantillon d'entrainement *i.i.d* $\mathcal{D}^L_n = \{(\mathcal{T}^{(1)}_T, Y^{(1)}), \ldots, (\mathcal{T}^{(n)}_T, Y^{(n)})\}$. A partir de $\mathcal{D}^L_n$, l'objectif est d'apprendre une règle de décision $\hat{g}$ qui prédit avec précision la classe à laquelle appartient une observation.

Présentons maintenant la procédure de classification ERMLR. Au préalable, on propose d'estimer p^* par sa contrepartie empirique $\hat{p}$. Ensuite, pour identifier les interactions effectives, une étape de sélection des interactions effectives, par classe, est conduite via l'estimateur Lasso présenté en section 2. S'ensuit, en tirant parti des support estimés par classe, d'une étape de réajustement basée sur un critère ERM (voir Vapnik (1991)).

Soit $\hat{\Theta}$ un ensemble d'intérêt, $\theta \in \hat{\Theta}$ et $f_{\hat{p},\theta}$ une fonction score associée. On considère le risque de classification L_2 suivant :

$$\hat{\mathcal{R}}_2\left(f_{\hat{p},\theta}\right) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(Z^{(i)}_k - f^k_{\hat{p},\theta}\left(\mathcal{T}^{(i)}_T\right)\right)^2, \quad Z^{(i)}_k = 2\mathbb{1}_{\{Y_i=k\}} - 1.$$

Ensuite, on définit l'estimateur de θ^* comme étant le minimiseur du risque empirique L_2

$$\hat{\theta}^R \in \underset{\theta\in\hat{\Theta}^R}{\arg\min}\, \hat{\mathcal{R}}_2(f_{\hat{p},\theta}).$$

où $\hat{\Theta}^R := \{\theta \in \hat{\Theta},\ \text{supp}(\theta_k) = \text{supp}(\hat{\theta}_k)\}$, avec $\text{supp}(\hat{\theta}_k)$ le support estimé de la k-ème classe donné par l'estimateur Lasso. Par conséquent, la minimisation est effectuée sur un ensemble de paramètres dont la dimension est beaucoup plus petite que d^2.

Nous pouvons énoncer le résultat suivant, on l'on énonce la consistance du classifier ERMLR. Ce dernier repose sur le résultat énoncé dans la section 2.

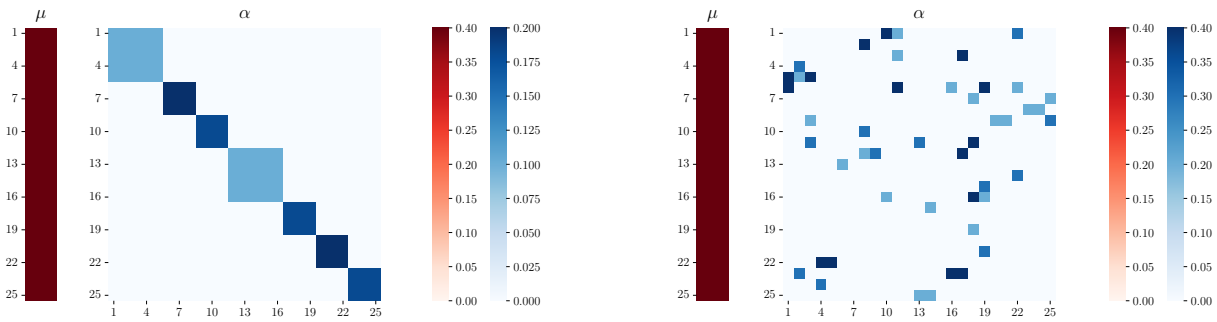
Théorème 2. *On suppose qu'il existe une constante* $s^* > 0$ *telle que* $\max_{k\in[K]}|\text{supp}(\theta_k^*)| \leq s^*$. *Pour* n *assez grand, il existe une constante* $C > 0$ *telle que,*

$$\mathbb{E}\left[\mathcal{R}(\hat{g}) - \mathcal{R}(g^*)\right] \leq C\left(\frac{(d+s^*)\log(nd)}{n}\right)^{1/2},$$

où C *dépend de* T, K *et d'autres constantes.*

4 Étude numérique avec le package Sparklen

On finit en proposant une étude numérique de la procédure sur données simulées. Cette dernière est conduite en s'appuyant sur la libraire Python **Sparklen** (Lacoste (2025)). **Sparklen** propose un ensemble complet d'outils de simulation, d'estimation et de classification au service de l'étude des PHM à noyau exponentiels. Nous considérons deux scénarios représentés visuellement dans la Figure 1.



(a) *Scénario 1* (b) *Scénario 2*

FIGURE 1 – Représentation de $\theta^* = (\mu^*, \alpha^*)$ dans les deux scénarios pour $d = 25$. Taux de parcimonie de α^* dans le *Scénario 1* : 85% ; dans le *Scénario 2* : 92%.

Pour illustrer la tâche de classification, nous considérons un problème de classification à $K = 3$ classes. Les trois classes sont créées sur la base des deux scénarios décrits ci-dessus. Les résultats sont donnés dans la table 1. Comme attendu, plus n est grand, plus le classificateur se rapproche du classificateur de Bayes, ce qui illustre la consistance de la procédure ERMLR.

	Scénario 1			*Scénario 2*		
	$n = 300$	$n = 600$	$n = 1500$	$n = 300$	$n = 600$	$n = 1500$
Bayes	0.087 (0.004)	0.086 (0.004)	0.087 (0.006)	0.237 (0.008)	0.236 (0.007)	0.236 (0.009)
ERMLR	0.134 (0.011)	0.113 (0.005)	0.098 (0.006)	0.326 (0.012)	0.276 (0.010)	0.254 (0.008)

TABLE 1 – Erreur empirique sur 30 répétitions Monte-Carlo dans les deux scénarios. L'impact de n est étudié. L'écart-type est indiqué entre parenthèses. $n_{\text{test}} = 3000$ est choisie. $T = 5$

5 Remerciements

Ce travail a été soutenu par la Chaire « Modélisation Mathématique et Biodiversité » de Veolia-École polytechnique-Museum national d'Histoire naturelle-Fondation X.

Références

C. Denis, C. Dion-Blanc, and L. Sansonnet. Multiclass classification for hawkes processes. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180, pages 539–547. PMLR, 2022.

C. Denis, C. Dion-Blanc, R. E. Lacoste, L. Sansonnet, and Y. Bas. Bats monitoring : A classification procedure of bats behaviors based on hawkes processes. *Journal of the Royal Statistical Society Series C : Applied Statistics*, 73(4) :1025–1041, 2024. doi : 10.1093/jrsssc/qlae024.

C. Dion-Blanc, C. Denis, L. Sansonnet, and R. E. Lacoste. Erm-lasso classification algorithm for multivariate hawkes processes paths. *arXiv preprint arXiv :2407.11455*, 2024.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

A. Hawkes. Cluster models for earthquakes-regional comparisons. *Bulletin of the International Statistical Institute*, 45(3) :454–461, 1973.

A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1) :83–90, 1971. doi : 10.1093/biomet/58.1.83.

R. E. Lacoste. Sparklen : A statistical learning toolkit for high-dimensional hawkes processes in python. 2025. URL `https://arxiv.org/abs/2502.18979`.

P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot. Inference of functional connectivity in neurosciences via Hawkes processes. In *2013 IEEE global conference on signal and information processing*, pages 317–320. IEEE, 2013.

V. Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991.

A subsampling scheme for the estimation of Hawkes processes through spectral theory

Felix Cheysson [1] & Miguel Martinez Herrera [2]

[1] *Gustave Eiffel University, France, felix.cheysson@univ-eiffel.fr*
[2] *Institut Pasteur, France, miguel.martinez-herrera@pasteur.fr*

Résumé. Les méthodes usuelles d'estimation pour les processus de Hawkes sont en général inaccessibles ou computationnellement coûteuses dans un contexte de données manquantes. Toutefois, avoir des observations partielles est un scénario commun lors de l'étude des données réelles, où le manque d'information peut être dû aux méthodes de détection d'évènements. Ce phénomène peut être modélisé mathématiquement par l'amincissement d'un processus ponctuel. Nous supposons que nous avons accès à une observation partielle d'un processus de Hawkes et nous proposons une procédure d'estimation paramétrique en maximisant une version de la log-vraisemblance issue de la théorie d'analyse spectrale. Nous proposons des conditions d'identifiabilité pour le modèle statistique qui en résulte pour le noyau exponentiel classique. Notre méthode d'inférence donne accès à une méthode de sous-échantillonage pour nos processus ponctuels à travers l'opération d'amincissement. Nous proposons ce paradigme comme outil pour améliorer la qualité de l'estimateur spectral pénalisé. Numériquement, nous montrons que dans un contexte avec un bas nombre d'observations, cet estimateur fournit un meilleur estimateur par rapport à d'autres approches.

Mots-clés. Processus de Hawkes, processus ponctuels, analyse spectrale, points manquants, sous-echantillonage.

Abstract. Standard estimation methods for Hawkes processes with missing information are either intractable or computationally expensive. Nonetheless this is a common scenario when studying real-world data where detection procedures may miss event times, which can be modelled through the thinning alteration of point processes. We assume that we observe a thinned version of a Hawkes process and we propose a parametric estimation procedure by maximising a version of the log-likelihood from spectral analysis theory. We propose identifiability conditions for the ensuing statistical model for the classical exponential kernel functions. Our inference method gives access to a subsampling method for our point processes via the thinning operation. We propose using this paradigm to enhance the performance of a penalised version of the spectral estimator. Numerically, we show that, in a small-sized sample context with a short observation window, this estimator performs better than alternative approaches.

Keywords. Hawkes processes, point processes, spectral analysis, missing points, subsampling.

1 Introduction

The Hawkes process model, introduced for the first time in Hawkes (1971), is a past-dependent point process used to characterise self-exciting interactions among event times. Its applications are numerous among various fields with examples including biology (Lambert et al., 2018), seismology (Ogata, 1988, 1998) and finance (Bacry et al., 2013; Hawkes, 2018), to mention a few.

As a result, there are many efforts focused around proposing estimation methods such as Ogata (1978); Ozaki (1979); Reynaud-Bouret et al. (2014); Bacry et al. (2020). These methods tend to study the model under the assumption that there is no missing data in the samples used for estimation. However, if this assumption does not hold, it is common for estimators to be biased.

The thinning operation on a point process can be used either to model missing data in point processes, to account for sporadic detection errors, or as a tool to provide a subsampling scheme for point processes (Møller and Waagepetersen, 2003; Cronie et al., 2024).

In this communication, we take inspiration from the spectral estimation methods as presented in Cheysson and Lang (2022); Bonnet et al. (2024). After a general presentation of the mathematical setting, we introduce the results concerning the spectral density function for thinned point processes. We leverage this result in the case of a Hawkes process with exponential kernel along with an identifiability condition for the statistical model. Finally, we illustrate the performance of our estimation procedure to enhance estimation through ℓ_2 penalisation with subsampling.

2 Mathematical setting

In this communication, we will focus on the stationary univariate Hawkes process on the real line $\mathbb{R}$, noted H. The Hawkes process can be characterised by its conditional intensity function defined, for all $t \in \mathbb{R}$, by:

$$\lambda(t) = \mu + \int_{-\infty}^{t} h(t-s)\, H(\mathrm{d}s) = \mu + \sum_{T_k \leq t} h(t - T_k)\,, \tag{1}$$

where $\mu > 0$ is the baseline intensity, and $h \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ is the kernel function, such that $\|h\|_1 = \int_{\mathbb{R}} h(t)\,\mathrm{d}t < 1$, modelling the self-exciting behaviour of past events $(T_k)_{k\in\mathbb{Z}}$.

In this communication, we will study a thinned version of process H defined as follows:

Definition 2.1. For any $p \in (0, 1]$, we denote H_p a p-thinning of a point process H defined, for any Borel set $B \in \mathcal{B}_{\mathbb{R}}$, as:

$$H_p(B) = \sum_{k\in\mathbb{Z}} \mathbb{1}_{T_k \in B} Z_k$$

where $(Z_k)_{k\in\mathbb{Z}}$ is an i.i.d. collection of Bernoulli random variables of parameter p. We refer to H_p as a p-thinned point process, and its event times correspond to a subset of $(T_k)_{k\in\mathbb{Z}}$ where each point is erased with a random probability $1 - p$.

3 Spectral model and estimation

3.1 Spectral theory of point processes

For any rapidly decreasing function $\varphi \in \mathcal{S}$, we define its Fourier transform $\tilde{\varphi}$, for any $\omega \in \mathbb{R}$, by: $\tilde{\varphi}(\omega) = \int_{\mathbb{R}} \varphi(x) \mathrm{e}^{-2\pi \mathrm{i} x\omega} \, \mathrm{d}x$. The Bartlett spectrum (Bartlett, 1963) of a second-order stationary point process H is the measure Γ such that, for any $\varphi \in \mathcal{S}$ (Brémaud et al., 2005, Definition 2):

$$\mathbb{V}\left[\int_{\mathbb{R}} \varphi(x) \, H(\mathrm{d}x)\right] = \int_{\mathbb{R}} |\tilde{\varphi}(\omega)|^2 \, \Gamma(\mathrm{d}\omega) \, . \tag{2}$$

Whenever Γ is an absolutely continous measure, we denote $f : \mathbb{R} \to \mathbb{C}$ its Radon-Nikodym derivative, known as the spectral density of process H.

For a realisation $(T_k)_{k=1:N(T)}$ of a process H in a time window $[0, T]$, the periodogram is defined, for all $\omega \in \mathbb{R}$, as:

$$I^T(\omega) = \frac{1}{T} \sum_{k=1}^{N(T)} \sum_{l=1}^{N(t)} \mathrm{e}^{-2\pi \mathrm{i} \omega (T_k - T_l)} \, .$$

In this communication, we will work on a parametric setting, so let a statistical model $\mathcal{P}$ for the spectral density be defined as $\mathcal{P} = \{f_\theta \colon \mathbb{R} \to \mathbb{C}, \theta \in \Theta\}$. We can then define the *spectral* log-likelihood $\ell_T(\theta)$ for an observation $(T_k)_{k=1:N(T)}$ of H as:

$$\ell_T(\theta) = -\frac{1}{T} \sum_{k=1}^{M} \left(\log(f_\theta(\omega_k)) + \frac{I^T(\omega_k)}{f_\theta(\omega_k)} \right) , \tag{3}$$

and then introduce the Whittle estimator $\hat{\theta}$ (Whittle, 1952) as $\hat{\theta} = \arg\max_{\theta \in \Theta} \ \ell_T(\theta)$.

3.2 Spectral density of a p-thinned point processes

As our model focuses on the thinning of a point process, the following general result can be first established.

Proposition 3.1. *Let H be a stationary point process admitting a the spectral density function f^H. For any $p \in (0, 1]$, let H_p be a p-thinning of H as explicited in Definition 2.1.*

Then, H_p admits a spectral density function, denoted f^{H_p}, such that for any $\omega \in \mathbb{R}$:

$$f^{H_p}(\omega) = p^2 f^H(\omega) + p(1-p) m_1 \, , \tag{4}$$

where $m_1 = \mathbb{E}[H([0, 1])]$ is usually known as the average intensity of process H.

For the rest of this communication, H will denote a stationary Hawkes process defined by the intensity λ as in Equation (1). As shown in Hawkes (1971), the spectral density of a

Hawkes process has an explicit expression. We focus on the exponential interaction function, defined as $\forall t \in \mathbb{R}$, $h(t) = \alpha\beta \mathrm{e}^{-\beta t}\mathbb{1}_{t\geq 0}$, where $\alpha \in (0,1)$ and $\beta > 0$. Equation (4) reduces to:

$$\forall \omega \in \mathbb{R}\,, \quad f_\theta(\omega) = \frac{\mu p}{1-\alpha}\left(1 + p\frac{\beta^2\alpha(2-\alpha)}{\beta^2(1-\alpha)^2 + 4\pi^2\omega^2}\right)\,.$$

In this context, the statistical model of f_θ is:

$$\mathcal{Q} = \{f_\theta \colon \mathbb{R} \to \mathbb{C}, \theta = (\mu, \alpha, \beta, p) \in \mathbb{R}_{>0} \times (0,1) \times \mathbb{R}_{>0} \times (0,1]\}\,.$$

and we obtain the following result concerning identifiability.

Proposition 3.2. *The model $\mathcal{Q}$ is not identifiable. In particular, for any admissible parameter $\theta = (\mu, \alpha, \beta, p)$ there exists an infinite number of parameters θ' such that $f_\theta = f_{\theta'}$.*

However, if we assume that one of the four quantities in θ is known, then the reduced model defined by the remaining triplet of parameters is identifiable.

In particular with this result, we can fix a parameter p of thinning and perform a thinning procedure on an observation of H to then obtain the subsampled Whittle estimator by maximising Equation (3).

4 Numerical results

In this section, we assume that we are provided with a single realisation H of a Hawkes process with parameters $(\mu, \alpha, \beta) = (1.25, 0.5, 1.5)$ in an observation window of $[0, 50]$.

To improve estimations in a context with few observations, we define then the ℓ_2-penalised spectral log-likelihood, for any penalisation parameter $L \geq 0$, as:

$$\ell_T(\theta) - L\|\theta\|_2\,. \tag{5}$$

A penalised estimator can then be obtained by maximising this quantity, and we will illustrate how combining this with a subsampling procedure by thinning provides better results than other approaches.

Benchmark estimators. We will compare four estimators:

$\hat{\theta}$: the estimator obtained by maximising the non-penalised spectral log-likelihood (Equation (3)) on the single observation H.

$\hat{\theta}^L$: the estimator obtained by maximising the penalised spectral log-likelihood (Equation (5)) on the single observation H.

$\hat{\theta}^L_{partition}$: we partition the observation window $[0, T]$ in I equally sized intervals. We average the penalised estimators obtained in each window to obtain $\hat{\theta}^L_{partition}$.

$\hat{\theta}^L_{thinning}$: we perform S different thinning procedures for a fixed $p \in (0,1)$ providing the p-thinned observations $H_1, \ldots, H_S$ of H. We average the penalised estimators obtained in each thinned observation to obtain $\hat{\theta}^L_{thinning}$.

Results. Table 1 displays for each estimator the MSRE of each parameter, the MSRE of θ and the proportion of times that each estimator outperforms the others. The hyper-parameters of each method are chosen such as to present the best estimators in the sense of the relative ℓ_2 error with respect to the true parameters of the model. We begin by observing that the non-penalised method provides by far the weakest estimator with a notably high MSRE when estimating β. On the other hand, the three penalised methods greatly improve the estimations, in particular for β, significantly reducing MSRE all around. We notice that the lowest MSRE are obtained by our thinning method with it providing, in 70% of the scenarios, the best estimations all around. This is a testimony of the advantages of the subsampling and penalisation procedure that we propose in this communication.

Estimator	MSRE				% best
	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\theta}$	
$\hat{\theta}$	0.18	0.13	5.14×10^2	2.85×10^2	1%
$\hat{\theta}^L$	0.12	0.09	0.13	0.12	6.7%
$\hat{\theta}^L_{partition}$	0.08	0.07	0.04	0.05	22.3%
$\hat{\theta}^L_{thinning}$	0.03	0.04	0.02	0.02	70%

Table 1: Mean square relative error (MSRE) of (μ, α, β) along with the MSRE of θ. Last column shows the proportion of times that each estimator achieves the lower relative ℓ_2 error accross all 1000 simulations.

5 Conclusion

In this communication, we leverage the spectral theory of point processes to present a parametric estimation procedure for thinned versions of univariate Hawkes processes. A first motivation for studying this model concerns the study of imperfect data. In particular, this procedure can make use of thinning as a tool to perform subsampling in scenarios with short observations of processes. By introducing a penalised version of the spectral log-likelihood and combining this with subsampling, we demonstrate that we are able to significantly improve estimations. This is encouraging for real-world applications, such as in health or biology, where acquiring sizeable quantities of data can be difficult.

The main extension of our work would be to adapt our methods to higher dimensional data in order to consider multiple inter-connected phenomena. Another avenue of exploration would be to propose a model selection procedure adapted to spectral approaches in order to choose the hyper-parameters of the penalised and subsampling methods in practical contexts where information on the real parameters is unavailable.

References

Bacry, E., Bompaire, M., Gaïffas, S., and Muzy, J. (2020), "Sparse and low-rank multivariate Hawkes processes," *Journal of Machine Learning Research*, 21, 1–32.

Bacry, E., Delattre, S., Hoffmann, M., and Muzy, J. (2013), "Scaling limits for Hawkes processes and application to financial statistics," *Stochastic Processes and Applications*, 123, 2475–2499.

Bartlett, M. S. (1963), "The Spectral Analysis of Point Processes," *Journal of the Royal Statistical Society. Series B (Methodological)*, 25, 264–296.

Bonnet, A., Cheysson, F., Martinez Herrera, M., and Sangnier, M. (2024), "Spectral analysis for noisy Hawkes processes inference," Preprint at `https://arxiv.org/abs/2405.12581`.

Brémaud, P., Massoulié, L., and Ridolfi, A. (2005), "Power spectra of random spike fields and related processes," *Advances in Applied Probability*, 37.

Cheysson, F. and Lang, G. (2022), "Spectral estimation of Hawkes processes from count data," *The Annals of Statistics*, 50, 1722 – 1746.

Cronie, O., Moradi, M., and Biscio, C. (2024), "A cross-validation-based statistical theory for point processes," *Biometrika*, 111, 625–641.

Hawkes, A. (1971), "Spectra of Some Self-Exciting and Mutually Exciting Point Processes," *Biometrika*, 58, 83–90.

— (2018), "Hawkes processes and their applications to finance: a review," *Quantitative Finance*, 18, 193–198.

Lambert, R., Tuleau-Malot, C., Bessaih, T., Rivoirard, V., Bouret, Y., Leresche, N., and Reynaud-Bouret, P. (2018), "Reconstructing the functional connectivity of multiple spike trains using Hawkes models," *Journal of Neuroscience Methods*, 297, 9–21.

Møller, J. and Waagepetersen, R. (2003), *Statistical Inference and Simulation for Spatial Point Processes*, New York, NY: Chapman and Hall/CRC, first edition.

Ogata, Y. (1978), "The asymptotic behaviour of maximum likelihood estimators for stationary point processes," *Annals of the Institute of Statistical Mathematics*, 30, 243–261.

— (1988), "Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes," *Journal of the American Statistical Association*, 83, 9–27.

— (1998), "Space-Time Point-Process Models for Earthquake Occurrences," *Annals of the Institute of Statistical Mathematics*, 50, 379–402.

Ozaki, T. (1979), "Maximum likelihood estimation of Hawkes' self-exciting point processes," *Annals of the Institute of Statistical Mathematics*, 31, 145–155.

Reynaud-Bouret, P., Rivoirard, V., Grammont, F., and Tuleau-Malot, C. (2014), "Goodness-of-Fit Tests and Nonparametric Adaptive Estimation for Spike Train Analysis," *The Journal of Mathematical Neuroscience*, 4, 3.

Whittle, P. (1952), "Some results in time series analysis," *Scandinavian Actuarial Journal*, 1952, 48–60.

Testing procedures based on maximum likelihood estimation for Marked Hawkes processes

Maya Sadeler Perrin[1] & Anna Bonnet[2] & Charlotte Dion-Blanc[3] & Adeline Leclercq Samson[4]

[1] *LJK, UMR 5224, Univ. Grenoble Alpes, Grenoble INP, 38000 Grenoble, France, maya.sadeler@univ-grenoble-alpes.fr*
[2] *LPSM, UMR 8001, Sorbonne Université, 75005 Paris, France, anna.bonnet@sorbonne-universite.fr*
[3] *LPSM, UMR 8001, Sorbonne Université, 75005 Paris, France, charlotte.dion_blanc@sorbonne-universite.fr*
[4] *LJK, UMR 5224, Univ. Grenoble Alpes, Grenoble INP, 38000 Grenoble, France, adeline.leclercq-samson@univ-grenoble-alpes.fr*

Résumé. Les processus de Hawkes sont des processus ponctuels auto-excitants permettant de capturer la dépendance temporelle entre événements successifs, contrairement aux processus de Poisson qui supposent l'indépendance des occurrences. Le processus de Hawkes marqué généralise ce cadre en associant à chaque événement une marque, représentant des informations additionnelles susceptibles d'influencer la dynamique temporelle du processus. Ces modèles sont largement utilisés dans divers domaines, tels que la finance, l'écologie ou l'analyse des réseaux sociaux, où les variables externes jouent un rôle clé dans la probabilité d'occurrence des événements futurs. La diversité des processus de Hawkes – incluant des variantes auto-excitantes, inhibitrices ou marquées – nécessite le développement de procédures de test robustes pour sélectionner les modèles les plus pertinents et éviter un sur-ajustement dû à l'introduction de covariables inutiles ou de dépendances temporelles superflues. Bien que plusieurs tests aient été proposés dans la littérature, ils restent souvent contraints par des hypothèses simplificatrices et un manque de validation empirique. Nous nous attachons donc à étudier le comportement de ces tests lorsqu'ils sont appliqués au-delà des cadres théoriques usuels qui garantissent leur validité. Si différentes procédures sont étudiées dans l'article original, nous nous concentrons, dans ce résumé, sur un aspect particulier: les tests permettant d'évaluer l'influence des covariables sur le processus. Ce point est essentiel pour l'analyse des processus de Hawkes marqués, car il permet une meilleure compréhension du rôle des informations transportées par la marque dans l'évolution du processus.

Mots-clés. Processus de Hawkes, Modélisation, Estimation, Test, Sélection de model

Abstract. Hawkes processes are self-exciting point processes that capture the temporal dependence between successive events, unlike Poisson processes, which assume independence of occurrence. The marked Hawkes process generalizes this framework by associating with each event a mark, representing additional information likely to influence the temporal dynamics of the process. These models are widely used in fields such as finance, ecology and social network analysis, where external variables play a key role in determining the probability of future events. The diversity of Hawkes processes - including self-exciting, inhibiting or

marked variants - requires the development of robust testing procedures to select the most relevant models and avoid over-fitting due to the introduction of unnecessary covariates or superfluous temporal dependencies. Although several tests have been proposed in the literature, they often remain constrained by simplifying assumptions and a lack of empirical validation. We therefore set out to study the behavior of these tests when applied beyond the usual theoretical frameworks that guarantee their validity. Among the various procedures studied in the original article, several tests are proposed to assess model quality and the relevance of covariates. In this summary, however, we focus on one particular aspect: tests to assess the influence of covariates through their impact on the mark associated with events. This point is essential for the analysis of marked Hawkes processes, as it enables a better understanding of the role of the information carried by the mark in the evolution of the process.

Keywords. Hawkes process, Modeling, Estimation, Testing, Model selection

1 Introduction

Les processus de Hawkes, introduits initialement dans l'article de Hawkes (1971), sont des processus ponctuels auto-excitants qui capturent les dépendances temporelles entre les événements, ce qui les explique leur utilisation dans divers domaines tel que les neurosciences, la finance, les sciences sociales... Une extension de ce modèle est le processus de Hawkes marqué, où chaque événement est associé à une marque qui peut influencer l'intensité des occurrences futures. La diversité des processus de Hawkes – incluant des variantes auto-excitantes, inhibitrices ou marquées – nécessite le développement de procédures de test robustes pour sélectionner les modèles les plus pertinents et éviter un sur-ajustement. Dans ce résumé extrait de l'article Bonnet et al. (2024), nous présentons le modèle de Hawkes ainsi que son extension qui est marquée, leur estimation par maximum de vraisemblance, un test statistique permettant d'évaluer l'impact des marques sur les évènements temporels, ainsi que quelques simulations.

2 Modèle

Les processus de Hawkes sont des processus temporel qui se définissent par une fonction d'intensité conditionnelle qui dépend de l'historique des événements et modélise les interactions temporelles en jeu. En notant $N(t) = (N^1(t), \dots, N^d(t))$ un vecteur de d processus ponctuels, on dira que ce vecteur est un processus de Hawkes exponentiel et multidimensionnel si, pour chaque composante i, l'intensité du $i-$ème processus est données, conditionnelement au passé du processus $N(t)$, par la formule suivante.

$$\lambda_\theta^i(t) = m_i + \sum_{j=1}^{d} \int_{(-\infty,t)} a_{ij} e^{-b_{ij}(t-s)} N^j(ds),$$

où m_i est le taux d'émission spontanée du processus N^i, a_{ij} mesure l'intensité de l'influence du processus N^j sur le processus N^i, et b_{ij} est un paramètre contrôlant la durée caractéristique cette intéraction. Cette fonction d'intensité reflète le fait que chaque processus possède une activité de base, avec une excitation supplémentaire induite par son propre historique et celui des autres composantes.

Une version étendue du modèle de Hawkes est le modèle de Hawkes marqué. Dans ce modèle, chaque événement est associé à une marque aléatoire, qui peut influencer l'intensité des interactions. Plus précisément, on considère un processus ponctuel marqué d-dimensionnel, où chaque événement se produit a un instant T_i, une étiquette m_i indiquant la composante dans laquelle l'évènement s'est produit, et une marque κ_i donnant une information au moment de l'évènement. On dira que ce processus est un processus de Hawkes exponentiel, marqué et multidimensionnel si l'intensité conditionnelle du i-ème processus est donnée par :

$$\lambda_\theta^i(t) = m_i + \sum_{j=1}^{d} \int_{(-\infty,t)\times\mathcal{E}} a_{ij} e^{-b_{ij}(t-s)} \phi_{ij}(\kappa) N^j(ds \times d\kappa),$$

où $\phi_{ij}(\kappa)$ est une fonction modulant la force d'interaction en fonction de la marque κ. Cette fonction permet à la marque d'amplifier ou d'atténuer l'interaction entre les événements de différentes composantes. Suivant une formulation paramétrique pour le modèle marqué, les fonctions ϕ_{ij} sont choisies en fonction d'un paramètre $\gamma_{ij} \in \mathbb{R}$, et nous notons donc $\phi_{ij} := \phi_{\gamma_{ij}}$. On considèrera souvent $\phi_{\gamma_{ij}}$ sous la forme $x \mapsto e^{\gamma_{ij} x}$ ou $x \mapsto x^{\gamma_{ij}}$.

Afin de garantir que ce modèle se présente bien comme une extension du modèle multidimensionnel non marqué, nous supposons que, pour $\gamma = 0$, la fonction ϕ_γ est constante et égale à 1. Sous ces hypothèses, le modèle de Hawkes marqué ou non marqué se présentent donc comme des modèles emboités : le modèle marqué est une extension du modèle non marqué, lui-même une extension du modèle de Poisson classique.

3 Estimation par maximum de vraisemblance

Nous nous concentrons sur l'estimation des paramètres du modèle en utilisant l'approche du maximum de vraisemblance. Pour les processus de Hawkes marqués et non marqués, la fonction de vraisemblance peut s'écrire en fonction de l'intensité :

$$\ell_T(\theta, \psi) := \sum_{j=1}^{d} \left(\int_0^T \int_{\mathcal{E}} \log(\lambda_\theta^j(s)) N^j(ds \times d\kappa) + \int_0^T \int_{\mathcal{E}} \log(f_\psi(\kappa)) N^j(ds \times d\kappa) - \int_0^T \lambda_\theta^j(s) ds \right).$$

L'estimateur du maximum de vraisemblance (EMV) est alors obtenu en maximisant la vraisemblance :

$$(\widehat{\theta}_T, \widehat{\psi}_T) \in \underset{(\theta,\psi)\in\Theta\times\mathbb{R}}{\operatorname{argmax}} \ \ell_T(\theta, \psi).$$

Cet estimateur est consistant en temps long pour le modèle non marqué, garantissant ainsi que les paramètres estimés convergent vers leurs vraies valeurs lorsque la durée d'observation du processus augmente.

4 Procedure de test

Après avoir estimé les paramètres par maximum de vraisemblance, nous effectuons un test d'adéquation du modèle. Cette approche, introduite pour la première fois par Reynaud-Bouret et al. (2014), repose sur le sous-échantillonnage des répétitions d'un processus de Hawkes à partir duquel est calculé une pvaleur. Pour améliorer la fiabilité de cette approche, nous présentons une version de ce test dans laquelle cette procédure est effectuée plusieurs fois sur des sous-échantillions tirés aléatoirement. Cette approche permet d'obtenir une distribution de pvaleur et de la comparer à une distribution uniforme sur $[0, 1]$.

Test 1 H_0 *:* $((T_k^{(r)})_k)_{1\leq r\leq n}$ *représente n réalisations i.i.d. d'un processus ponctuel d'intensité* λ_{θ^*} *sur* $[0, T_{\max}]$*, où* θ^* *est un paramètre inconnu.*

1. *Fixer* $p(n) \in \{1, ..., n\}$ *tel que* $p(n)/n \xrightarrow[n\to\infty]{} 0$ *;*
2. *Pour chaque* $r \in \{1, ..., n\}$ *:*
   - *Calculer le EMV* $\hat{\theta}^{(r)}$ *associé à la réalisation* $(T_k^{(r)})$ *;*
   - *Calculer* $\hat{\theta}$*, la moyenne des* $(\hat{\theta}^{(r)})_r$ *;*
3. *Pour tout sous-ensemble* S *de* $\{1, ..., n\}$ *de cardinalité* $p(n)$ *:*
   (a) *Pour tout* $r \in S$*, calculer* $\hat{\mathcal{N}}^r := \{\hat{\Lambda}_i(T_k^{(r)}), 1 \leq k \leq \mathcal{N}^r_{T_{\max}}\}$*, avec* $\hat{\Lambda}_r(s) = \int_{(0,t)} (\lambda_{\hat{\theta}})_r(s)ds$ *et* $(\lambda_{\hat{\theta}})_r$ *est l'intensité d'un processus de Hawkes de paramètre* $\hat{\theta}$ *associé aux temps d'arrivée* $(T_k^{(r)})$ *;*
   (b) *Construire* $\hat{\mathcal{N}}^{cum}$*, le processus cumulé associé à* $(\hat{\mathcal{N}}^r)_r$*, par récurrence sur les éléments de* $S := \{r_1, ..., r_{p(n)}\}$ *:*
      - ***Initialisation,*** $j = 1$ ***:*** $\hat{\mathcal{N}}^{cum} \leftarrow \hat{\mathcal{N}}^{r_1}$ *;*
      - ***Récurrence*** *pour* $1 \leq j \leq p(n)$ *:* $\hat{\mathcal{N}}^{cum} \leftarrow \hat{\mathcal{N}}^{cum} \cup \left(\sum_{l<j} \hat{\Lambda}_{r_l}(T_{\max}) + \hat{\mathcal{N}}^{r_j}\right)$ *;*
   (c) *Choisir* $\xi > 0$ *tel que* $\xi p(n) < \sum_{r\in S} \hat{\Lambda}_r(T_{\max})$ *presque sûrement ;*
   (d) *Effectuer un test pour comparer la distribution du processus cumulé* $\left(\frac{1}{\xi p(n)}\hat{\mathcal{N}}^{cum}_t\right)_{t\in[0,\xi p(n)]}$ *à une distribution uniforme sur* $[0, 1]$ *et récupérer la p-valeur associée* p_S *;*
4. *Comparer* $(p_S)_S$ *à une distribution uniforme.*

Grâce à cette procédure, il est possible de tester l'adéquation du modèle de processus

de Hawkes aux données, offrant ainsi un critère de sélection pour les modèles disponibles. Bien que la vraie valeur du paramètre θ^* n'intervienne pas directement dans le test, elle joue un rôle implicite car, pour le processus $\frac{1}{\xi p(n)}\hat{\mathcal{N}}^{\text{cum}}$ converge vers une distribution uniforme, l'estimateur de paramètre $\hat{\theta}$ doit vérifier la condition de convergence en probabilité suivante:

$$p(n)^{-1/2}\left[\sum_{i\in S}\left\|\left(\lambda_{\hat{\theta}}\right)_i-\left(\lambda_{\theta^*}\right)_i\right\|_{l_1([0;T_{\max}])}\right]\xrightarrow[n\to\infty]{\mathbb{P}}0.$$

Ainsi, les seules hypothèses pour ce test concernent le type de modèle choisi et la qualité de l'estimation réalisée sous l'hypothèse nulle, comme l'indique la convergence en probabilité mentionnée.

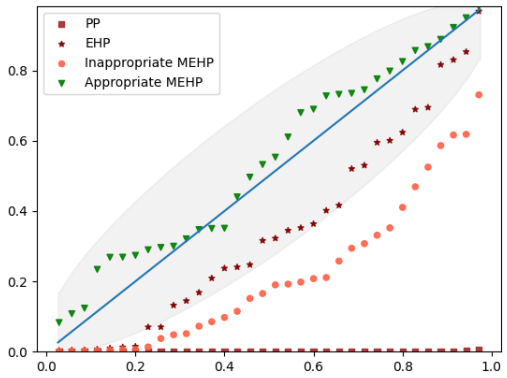
5 Exemple de simulation

Comme dit précédement, ce test repose sur un certain contrôle de la norme l_1 de l'estimateur de l'intensité. Comme cette condition n'a pas été vérifiée pour l'estimateur EMV, dans le cadre des modèles de Hawkes, nous présentons des résultats d'une étude numérique visant à vérifier la performance du test dans ce contexte de Hawkes marqué. Le Test 1 nous amenant à comparer des p-valeurs à une distribution uniforme, nous les représentons dans un graphe de type qqconf, basé sur la librairie `qqconf` de R qui permet de construire des intervalles de confiance adaptés autour de la distribution théoriques testés.

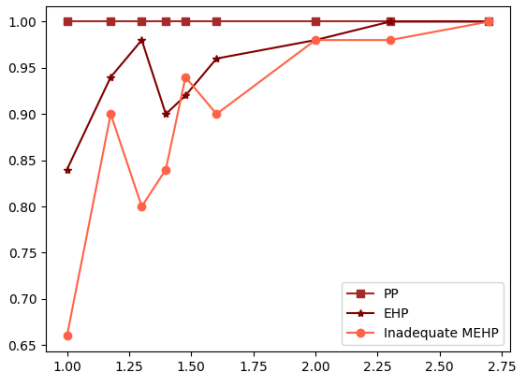
La Figure 1a montre que le test permet bien de différencier les différents modèles envisagés. En effet, le seul modèle accepté (car ayant des pvaleur dont on ne rejette pas l'uniforme distribution) est celui correspondant au modèle utilisé lors des simulations, représenté par la courbe verte. Par ailleurs, on observe une amélioration des puissances lorsque le nombre de répétitions augmente, ce qui est attendu. Toutefois, une exception notable concerne le test associé au processus de Poisson, qui rejette immédiatement l'hypothèse d'absence de dépendance temporelle, indépendamment du nombre de répétitions.

6 Conclusion

Dans ce travail, nous présentons le modèle de Hawkes marqué ainsi qu'une procédure de test permettant d'évaluer sa pertinence, accompagnée d'une validation empirique lorsque le cadre théorique garantissant sa validité n'est plus applicable. L'article de référence explore également d'autres procédures de test, complétées par une étude de simulation approfondie visant à identifier, empiriquement, les conditions cas dans lesquelles ils peuvent s'appliquer, offrant ainsi des outils complémentaires pour affiner l'évaluation et le choix du modèle le plus adapté.



(a) Test d'uniformité pour les p-valeurs du Test 1.



(b) Puissance du test en fonction du logarithme du nombre de répétitions.

Figure 1: Test d'uniformité des p-valeurs du Test 1 (Figure 1a) et puissance du Test (Figure 1b) lors de la simulation d'un processus de Hawkes marqué avec les paramètres $m^\star = 1$, $a^\star = 0.6$, $b^\star = 1$, $\phi_{\gamma,\psi}(x) = \frac{\psi-\gamma}{\gamma} e^{x\gamma}$ avec $\gamma^\star = 0.5$, et $(\kappa_i)_i$ i.i.d. suivant une loi exponentielle de paramètre $\psi^\star = 1$. Les simulations incluent un test pour le Poisson (PP), pour le Hawkes exponentiel non marqué (EHP), pour le Hawkes marqué avec $\phi_{\gamma,\psi}(x) = \frac{\psi-\gamma}{\gamma} e^{x\gamma}$ (Appropriate MEHP) et un modèle de Hawkes marqué avec $\phi_{\gamma,\psi}(x) = \psi^\gamma \Gamma(\gamma+1)^{-1} x^\gamma$ (Inappropriate MEHP).

Bibliographie

Reynaud-Bouret, P., Rivoirard, V., Grammont, F., Tuleau-Malot, C. (2014), Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis *The Journal of Mathematical Neuroscience*, 4, pp. 1-47.

Hawkes, Alan G. (1971), Point spectra of some mutually exciting point processes, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 33, pp. 438-443.

Bonnet, A. , Dion-Blanc, C., and Perrin Sadeler, M. (2024),Testing procedures based on maximum likelihood estimation for Marked Hawkes processes, *arXiv preprint arXiv:2410.05008*.

Statistique et sport 1

Modèle prédictif pour les performances de course à pied avec données fonctionnelles scalaires et qualitatives.

François Bassac[1,2] & Cristian Preda[2] & Cédric Morio[1]

[1] *Decathlon Lille, France, francois.bassac@decathlon.fr, cedric.morio@decathlon.com*
[2] *MODAL Inria Lille , France, cristian.preda@univ-lille,fr*

Résumé. Prédire les performances des coureurs utilisant des programmes d'entraînement proposés par Decathlon est un défi majeur pour l'entreprise mais aussi pour les coureurs. Des données mesurant les capacités physiques du coureur, la thématique de l'entraînement suivi et les issues de séances d'entraînement sont disponibles chez Decathlon. Leur caractère longitudinal (dimension temporelle) et leur hétérogénéité (type, horizon d'observation) nécessitent un prétraitement performant pour leur utilisation dans un modèle prédictif. L'extension de l'analyse en composantes principales et de l'analyse des correspondances multiples aux données fonctionnelles scalaires, respectivement qualitatives, est utilisée pour réduire la dimension, visualiser les données et ajuster un modèle de régression linéaire. Les fonctions coefficients du modèle de régression permettent l'interprétation et la prédiction avec des nouvelles données.
Une application numérique sur un échantillon de 1120 courses à pied illustre la méthodologie proposée.

Mots-clés. données fonctionnelles qualitatives, analyse en composantes principales, analyse des correspondances multiples, régression, prédiction.

Abstract. Predicting the performance of runners using the training programs offered by Decathlon is a major challenge for both the company and the runners themselves. Data measuring the runners physical capacities, the training themes followed and the results of training sessions are available at Decathlon. Their longitudinal nature (time dimensionality) and heterogeneity (type, length of observation) require effective pre-processing before they can be used in a predictive model. The extension of principal component analysis and multiple correspondence analysis to scalar and categorical functional data is used to reduce the dimension, visualize the data and fit a linear regression model. The coefficient functions of the regression model allow interpretation and prediction with new data.
A numerical application on a sample of 1120 running events illustrates the proposed methodology.

Keywords. categorical functional data, principal component analysis, multiple correspondence analysis, regression, prediction.

1 Introduction

KIPRUN PACER est une des applications sportives de Decathlon dédiée à la course à pied. Elle propose des plans d'entraînement adaptés aux coureurs et coureuses en fonction de leurs capacités sportives et de leur objectif à une course donnée (10 kilomètres, semi-marathon ou marathon). À l'issue de la réalisation du plan d'entraînement par le coureur, un des objectifs de l'application KIPRUN PACER est de prédire la performance effective du coureur à l'épreuve pour laquelle il s'est entrainé. Cette prédiction permettra aussi de savoir si le temps cible (ambition du coureur) est atteignable ou non. Certaines études sportives ont permis d'apporter des modèles simples de cette performances future en fonction des caractéristiques sportives des utilisateurs résumées à leur *Vitesse Maximale Aérobie* (VMA) comme le modèle de Riegel (1981). Dans ce travail nous proposons d'intégrer dans un même modèle de régression les données caractérisant le coureur (VMA, allure, temps visé), la course (distance courue) et celles issues des séances d'entraînement. Un plan d'entraînement se déroule sur plusieurs semaines avec une ou plusieurs séances d'entraînement par semaine. Pour chaque séance, on dispose de la thématique de l'entraînement (propre à chaque coureur) et de la qualité de la réalisation de l'entraînement par celui-ci. La thématique de l'entraînement (qui peut changer d'une

séance à l'autre) est une donnée qualitatives à six états : *Test VMA, Optimisation de la VMA, Vitesse spécifique, Endurance, Développement de l'endurance, Affutage*. La qualité de l'entraînement quant à elle est un score (note) à valeurs dans l'intervalle $[0, 1]$, avec zero correspondant à une séance non-effectuée. Un exemple de plan d'entraînement et de la qualité de sa réalisation est donnée dans la Figure 1.

D'autres modélisations sportives plus récentes existent comment celles de Mulligan (2018) ou d'Emig & Peltonen (2020). Mais leur précision fine pour des coureurs "élites" et non récréationnels, l'utilisation de données très précises de courses passées réalisées au maximum des capacités du sportif, ainsi que leur équation plus compliquée à mettre en oeuvre (appel à la fonction W de Lambert, minimisation à deux ou trois inconnues) n'en font pas des modèles aisément applicables au plus grand nombre de coureurs.

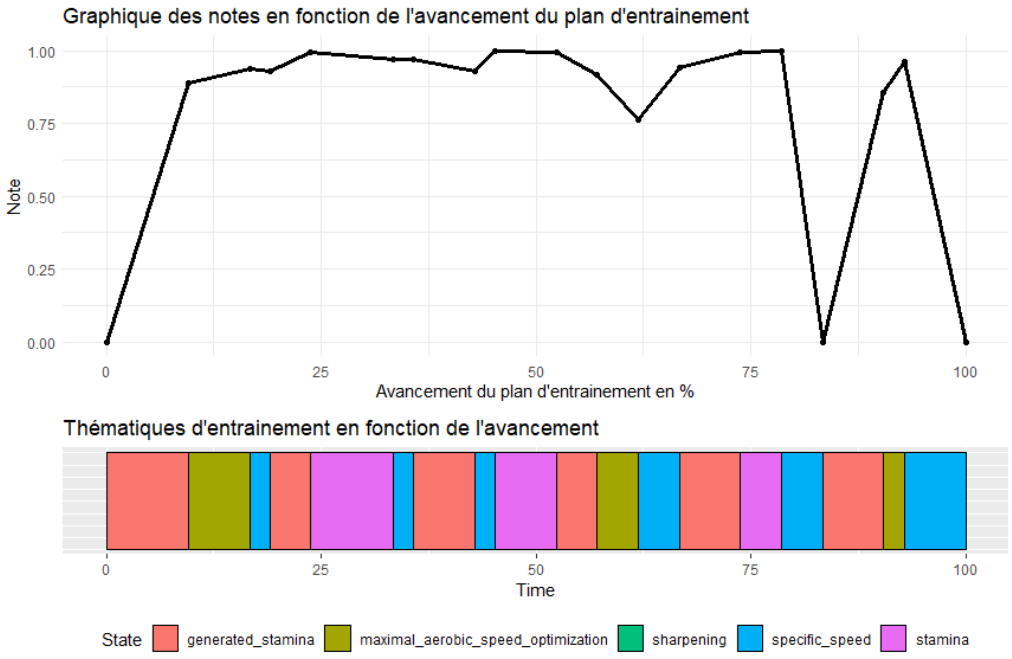


FIGURE 1 – Thématiques d'un plan d'entraînement (bas) et sa réalisation (haut)

2 Modèle de régression avec prédicteurs fonctionnels scalaires et qualitatifs

Dans cette modélisation nous nous intéressons à la prédiction de la performance des coureurs après que le plan d'entrainement soit entièrerment effectué par ceux ci.

La performance réalisée par un coureur après un plan d'entraînement est une variable aléatoire à valeurs réelles notée Y. Les changements des thématiques d'entraînement et l'évolution de la qualité des réalisations des séances (Figure 1) sont considérées comme des données de type fonctionnel, chaque trajectoire étant vue comme la réalisation d'un processus stochastique à temps continu. Les changements des thématiques d'un entraînement sont associés à un processus qualitatif générant des données fonctionnelles qualitatives (Preda et. al (2021)) alors que les scores de réalisation des séances d'entraînement ont comme modèle sous-jacent un processus scalaire (Ramsay and Silverman (2007)).

Les prédicteurs considérés dans le modèle de régression sont :

— $X_1 = (X_{1,1}, \ldots, X_{1,p})$: vecteur aléatoire à valeurs dans $\mathbb{R}^p$, $p \geq 1$, regroupant les données du coureur (VMA, allure, temps visé), de course (distance courue) et nombre de jours d'entraînement.

— $X_2 = \{X_2(t), t \in [0,T]\}$: variable aléatoire fonctionnelle scalaire représentant l'évolution sur la période d'entraînement [0,T] de la qualité de la réalisation des séances d'entraînement, $X_2(t)$ est un score dans l'intervalle $[0,1]$ (note normalisée) mesurant la qualité de la réalisation de la séance prévue au temps t (exprimé en pourcentage du temps d'entraînement, T = 100%).
— $X_3 = \{X_3(t), t \in [0,T]\}$: variable aléatoire fonctionnelle qualitative décrivant la succession des thématiques d'entraînement : pour une séance réalisée au temps t, $X_3(t)$ est une variable qualitative avec $K = 6$ modalités (états) : *Test VMA, Optimisation de la VMA, Vitesse spécifique, Endurance, Développement de l'endurance, Affutage*. La variable $X_3(t)$ est résumée à l'aide des fonctions indicatrices de chaque état, $X_3(t) = \{\mathbb{1}_t^s, s = 1, \ldots, K\}$.

Nous proposons le modèle de régression suivant :

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_{1,i} + \int_0^T \beta(t) X_2(t) \mathrm{d}t + \int_0^T \sum_{s=1}^{K} \gamma_s(t) \mathbb{1}_t^s \mathrm{d}t + \epsilon, \tag{1}$$

où $\{\beta_i\}_{i=0,p}$, $\beta : [0:T] \to \mathbb{R}$ et $\gamma_s : [0:T] \to \mathbb{R}$, $s \in 1, \ldots, K$, représentent les coefficients du modèle.

Estimation. Afin d'estimer les paramètres fonctionnels β et $\{\gamma_s\}_{s=1,\ldots K}$, les trajectoires résumant la qualité de l'entrainement (courbe scalaire de notes) et celle de la thématique d'entrainement (courbe qualitative) sont résumées dans le modèle à travers les composantes principales des deux analyses de données fonctionnelles. Les variables fonctionnelles sont alors représentées à l'aide des composantes principales $\{z_i\}_{i\geq 1}$:

- pour la variable fonctionnelle scalaire (Ramsay and Silverman (2005)) :

$$X_2(t) = \mu(t) + \sum_{i\geq 1} z_i u_i(t), \quad t \in [0,T],$$

où μ et $\{u_i\}_{i\geq 1}$ sont la moyenne, respectivement les fonctions propres de l'opérateur de covariance, du processus X_2.
- pour la variable fonctionnelle qualitative (Preda et al (2021)),

$$\mathbb{1}_t^s = p_s(t) + \sum_{i\geq 1} z_i a_i^s(t) p_s(t), \quad t \in [0,T]$$

où $p_s(t) = \mathbb{P}(X_3(t) = s)$ et $\{a_i^s\}_{i\geq 1, s\in 1:K}$ sont les fonctions optimales de codage du processus X_3 associées aux composantes principales $\{z_i\}_{i\geq 1}$.

Le calcul des composantes principales est réalisé à l'aide d'une approximation des facteurs principaux $\{u\}_{i\geq 1}$ et des fonctions de codage $\{a_i^s\}_{i\geq 1, s\in 1:K}$ dans une base de fonctions splines cubiques (B-Splines). Une procédure de sélection des composantes principales basée sur le critère AIC est appliqué à l'ensemble des premières composantes assurant la conservation d'au moins 90% de l'information (variance totale).

3 Application

Le modèle (1) est estimé à partir d'un échantillon de 1120 courses à pied pour lesquelles les données sont disponibles chez Decathlon.

Les coefficients de variables non-fonctionnelles (X_1) sont présentés dans la Table 1.

Les fonctions de régression β et $\{\gamma_s\}_{s\in 1:K}$ sont présentées dans la Figure 2.

Nous comparons les résultats de ce modèle avec le modèle de Riegel qui fait référence dans le domaine de la course à pied (Riegel (1981)). Dans ce dernier modèle, seule la variable VMA du coureur compte (quantité prenant en compte les données historiques du coureur). On obtient les résultats suivants :

L'indicateur de performance RMSEP (en secondes) présenté Table 2 est calculé par validation croisée (*leave-one-out*). L'écart-type de l'erreur de prédiction pour le modèle intégrant les co-variables fonctionnelles

Variable	Coefficient	p-value
VMA (km/h)	-215.83195	3.57e-05
Allure de course (s/km)	-13.24704	9.89e-08
Distance (m)	-0.05484	0.0207
Jours d'entrainements	-3.76544	0.0872
Temps cible	1.24860	$<$ 2e-16
Intercept	7315.33569	7.89e-08

TABLE 1 – Coefficients des variables non-fonctionnelles

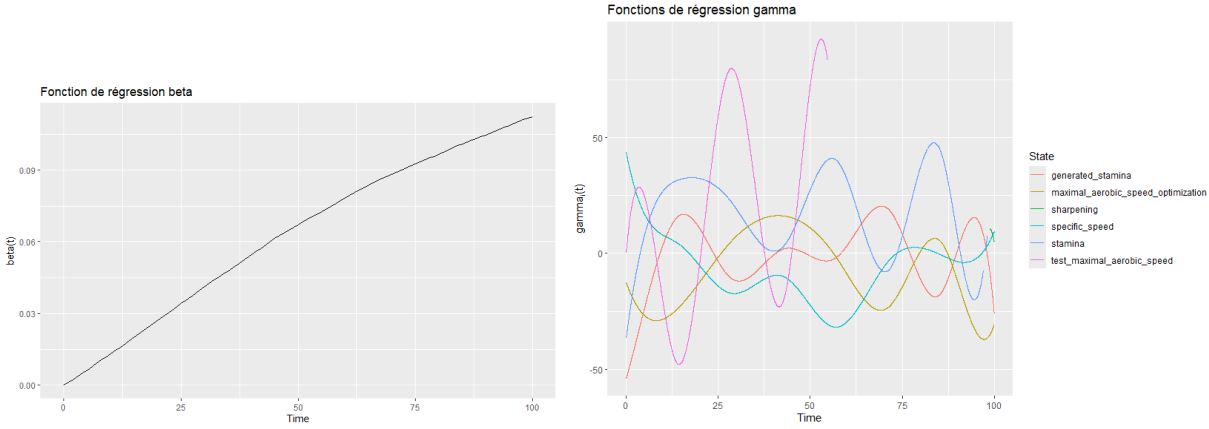


FIGURE 2 – Coefficients de régression fonctionnels : β (gauche) et γ_s (droite)

comme prédicteurs est de l'ordre de 25 minutes. Ce modèle apporte une amélioration de l'ordre de 16% sur ce même écart-type par rapport au modède de Riegel.

La Figure 3 présente une visualisation des données prédites et réalisées et la Figure 4 présente les résidus.

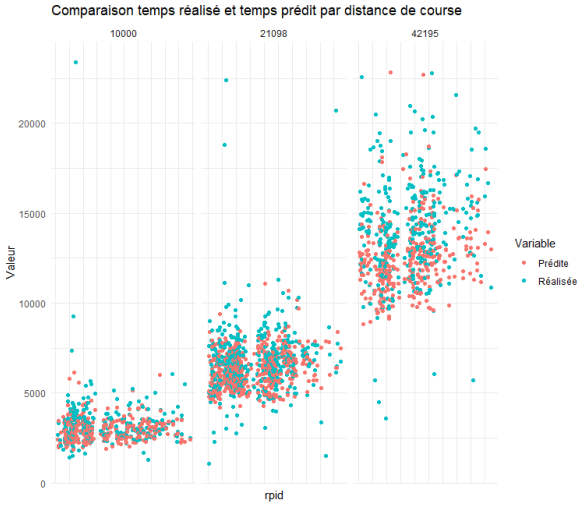


FIGURE 3 – Plot comparatifs des temps réalisés, temps prédits

La modélisation présentée performe mieux que le modèle de Riegel sur l'ensemble des courses. L'unicité du modèle pour toutes les distances est une contraite de modélisation posée par Kiprun PACER. Des modèles dédiés à chaque distance pourraient être plus performants que le modèle multi-courses.

	Modèle	Modèle de Riegel
RMSEP	1.534254e+03	1.831826e+03

TABLE 2 – Performances de prédiction

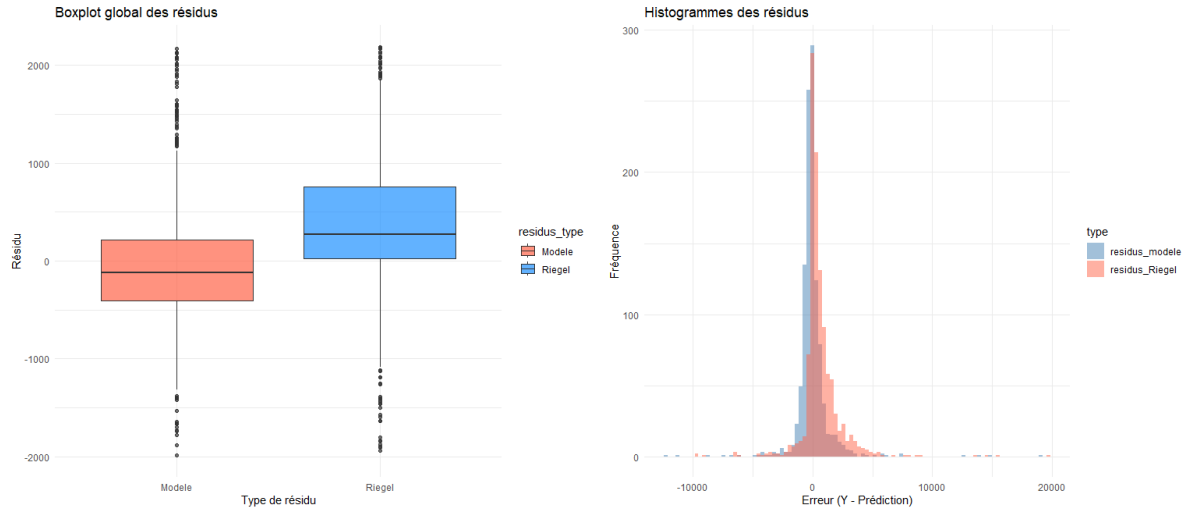


FIGURE 4 – Boxplot et histogramme des résidus.

La Table 3 présente l'indice de Shapley résumant l'importance des variables dans le modèle linéaire estimé. Notons la contribution importante de la variable *temps cible* qui confirme la connaissance par le coureur des ses propres capacités. Malgré une faible contribution des variables fonctionnelles (évaluation de l'entrainement et choix du plan d'entrainement), celles-ci s'avèrent néanmoins significatives dans la modélisation (test ANOVA des modèles emboités).

4 Conclusions et perspectives

La méthode présentée permet d'exploiter dans un même modèle des données scalaires non fonctionnelles, des données fonctionnelles scalaires et catégorielles. Les méthodes d'analyse de données fonctionnelles permettent à l'aide des composantes principales de réduire la dimension des prédicteurs fonctionnels et ainsi d'estimer une vaste gamme de modèles de régression.

Dans l'application de la prédiction de performance en course à pied, cette approche permet de diminuer de 16,3% l'écart-type de l'erreur de prédiction et de prendre en compte à la fois les thématiques d'entraînement d'un coureur ainsi que ses réalisations.

Dans le cadre d'une plus large utilisation de cette méthodologie, nous nous intéressons à son extension à une approche de prédiction "online".

Bibliographie

Preda, C., Grimonprez, Q., & Vandewalle, V. (2021). *Categorical Functional Data Analysis. The cfda R Package.* Mathematics, 9(23), 3074. https ://doi.org/10.3390/math9233074

Ramsay, J.O. and Silverman, B.W. (2005) Functional Data Analysis. Springer, New York.
https ://doi.org/10.1002/0470013192.bsa239

Variable	Valeur	Valeur standardisée
Temps cible	0.3877	0.4327
Distance (m)	0.3441	0.3840
Rythme de course (s/km)	0.0543	0.0606
Jours d'entrainements	0.0433	0.0483
VMA (km/h)	0.0202	0.0226
FMCA cp 5	0.0113	0.0126
FMCA cp 7	0.0102	0.0114
FMCA cp 9	0.0085	0.0095
FMCA cp 19	0.0048	0.0053
FMCA cp 33	0.0040	0.0045
FMCA cp 23	0.0038	0.0043
FMCA cp 1	0.0017	0.0019
FPCA cp 1	0.0010	0.0011
FMCA cp 21	6e-04	7e-04
FMCA cp 30	4e-04	4e-04

TABLE 3 – Valeurs ordonnées de Shapley du modèle

Deville, J.C. ; Saporta, G. Correspondence Analysis with an Extension towards Nominal Time Series. J. Econom. 1983, 22, 169–189.

Ramsay, J.O. ; Wickham, H. ; Graves, S. ; Hooker, G. fda : Functional Data Analysis, R Package Version 5.5.1 ; 2021. Available online : https ://cran.r-project.org/package=fda (accessed on 28 November 2021).

G.Saporta, Méthodes exploratoires d'analyse de données temporelles, Cahiers du BURO, 37–38, 1981.

Aguilera, A., Escabias, M., Preda, C., Saporta, G. (2010), "Using basis expansions for estimating functional PLS regression : Applications with chemometric data," Chemometrics and Intelligent Laboratory Systems, 104, 289-305.

Preda, C.,Saporta, G. (2005), "PLS regression on a stochastic process," Computational Statistics and Data Analysis, 48, 149-158.

Riegel, Peter S. "Athletic Records and Human Endurance : A Time-vs.-Distance Equation Describing World-Record Performances May Be Used to Compare the Relative Endurance Capabilities of Various Groups of People." American Scientist 69, no. 3 (1981) : 285–90. http ://www.jstor.org/stable/27850427.

Mulligan M, Adam G, Emig T (2018) A minimal power model for human running performance. PLoS ONE 13(11) : e0206645. https ://doi.org/10.1371/journal.pone.0206645

Emig, T., Peltonen, J. Human running performance from real-world big data. Nat Commun 11, 4936 (2020). https ://doi.org/10.1038/s41467-020-18737-6

CARACTÉRISATION ET ÉTUDE DES PROFILS PSYCHOLOGIQUES EXISTANTS DANS UN ÉCHANTILLON DE TRAILEURS

Clément Baud[1], Alice Chauveau[2], Franck Corset[3], Stéphane Doutreleau[4], Clément Ginoux[5], Adeline Samson Leclercq[6]

[1] *Laboratoire SENS et HP2, Université Grenoble Alpes, Inserm, France, clement.baud@univ-grenoble-alpes.fr*
[2] *CDP SportPerfHealth, GRICAD, Université Grenoble Alpes, France, alice.chauveau@univ-grenoble-alpes.fr*
[3] *LJK et IUT2, Université Grenoble Alpes, France, franck.corset@univ-grenoble-alpes.fr*
[4] *UM Sports et Pathologies, Clinique Physiologie Sommeil Exercice, Pôle Thorax et Vaisseaux, Hôpital Sud, CHU de Grenoble, Laboratoire HP2, Inserm, France, sdoutreleau@chu-grenoble.fr*
[5] *Laboratoire SENS, Université Grenoble Alpes, France, clement.ginoux@univ-grenoble-alpes.fr*
[6] *LJK, Université Grenoble Alpes, France, adeline.leclercq-samson@univ-grenoble-alpes.fr*

Résumé. Le trail est une activité sportive pratiquée par de plus en plus de personnes. Cette pratique consiste à effectuer un parcours de course à pied dans un environnement naturel sur des distances allant de quelques kilomètres à plusieurs centaines. Toutes ces spécificités engendrent des contraintes physiologiques (e.g., temps ou intensité d'effort important), psychologiques (e.g., réussir à maintenir sa motivation, son engagement dans l'activité) et cognitives (e.g., gestion de la fatigue mentale, de la perte de lucidité) importantes. Au regard des spécificités de la pratique, il est naturel de se questionner sur le profil psychologique (i.e., ensemble de caractères psychologiques) des individus s'impliquant dans cette discipline. L'utilisation de méthode de clusterisation permet de faire ressortir les profils psychologiques existants dans cette population. De plus, cette analyse permettrait d'évaluer les éventuels liens entre ces profils psychologiques et des variables comportementales (i.e., blessure, addiction, performance). Ces informations pourront être utiles pour les pratiquants de cette activité et parallèlement aux profesionnels intervenants dans celle-ci (i.e., entraîneurs, médecins).

Mots-clés. Sport, Trail, profil psychologique, clustering, modèle de mélange, analyse factorielle confirmatoire.

Abstract. Trail running is an activity practised by more and more people. It involves running in a natural environment over distances ranging from a few kilometres to several hundred. All of these specificities generate significant physiological (e.g. long duration or intensity of effort), psychological (e.g. managing to maintain motivation and commitment to the activity) and cognitive (e.g. managing mental fatigue and loss of lucidity) constraints. Given the specific nature of the sport, it is natural to ask questions about the psychological profile (i.e. a set of psychological characteristics) of individuals involved in this discipline. The use of clustering methods can help to highlight the psychological profiles that exist in this population. In addition, this analysis will make it possible to assess possible links between these psychological profiles and behavioural variables (i.e. injury, addiction, performance). This information could be useful for those practising this activity and also for the professionals involved in it (i.e. trainers, doctors).

Keywords. Trail, psychological profil, clustering, mixture model, confirmatory factor analysis.

1 Contexte

Le trail est une pratique consistant à effectuer un parcours de course à pied ouvert à tous qui se déroule dans un environnement naturel (International Trail Running Association[3]).

Nous observons une nette augmentation du nombre de traileurs et d'évènements de trail depuis ces dernières années. Cette émergence est marquée par la pratique, en France, de près d'un million de trailers en 2018, accompagnés d'un nombre de compétitions toujours plus important chaque année, jusqu'à plus de 4500 cette même année (Plard & Martineau, 2021[5]).

Le trail est une activité pratiquée sur des distances allant de quelques kilomètres à plusieurs centaines de kilomètres sur des parcours incluant de plus ou moins longues portions de montées et de descentes le tout dans un environnement naturel (e.g., montagne, bord de mer, désert) avec ses propres caractéristiques. Toutes ces spécificités engendrent des contraintes physiologiques (e.g., temps ou intensité d'effort important), psychologiques (e.g., réussir à maintenir sa motivation, son engagement dans l'activité) et cognitives (e.g., gestion de la fatigue mentale, de la perte de lucidité) importantes. Il n'est pas rare d'entendre des remarques comme : « est-ce qu'ils sont fous ? » ; « pourquoi font-ils ça ? » quand on évoque la pratique de ce sport. Au regard des spécificités de la pratique, il est naturel de se questionner sur les variables psychologiques propres aux individus s'impliquant dans cette discipline.

Si certaines études se sont intéressées aux différentes variables psychologiques propres à cette population (Roebuck et al., 2018[7]), à l'heure actuelle peu d'études abordent cette question par le prisme des profils psychologiques (i.e., ensemble de caractères psychologiques). Or cette approche permettrait de mieux caractériser psychologiquement ces individus ainsi que les éventuels liens entre les variables psychologiques et comportementales (i.e., blessure, performance).

L'objectif de cette étude est d'analyser statistiquement une étude réalisée sur des traileurs et d'identifier par clustering des groupes d'individus au profil psychologique similaire.

2 Méthodes

2.1 Participants et procédure

L'Ultra-tour des 4 Massifs (Ut4M) est un événement de trail de la région grenobloise regroupant 12 courses. Dans le cadre du projet interdisciplinaire Sport Perf Health de l'Université Grenoble Alpes, une collaboration est déjà effective entre l'Association Grenoble Outdoor Aventure, organisatrice de l'Ut4M, et plusieurs équipes de recherche académiques de l'Université Grenoble Alpes.

Lors des éditions 2023 et 2024 de l'Ut4M, un appel à volontaires a été effectué via leur site internet. Un questionnaire en ligne a été diffusé à tous les participant.e.s par le biais de l'organisation. Les données recueillies sont issues de la concaténation des questionnaires de 2023 et 2024. L'échantillon exploitable pour ces analyses comporte les réponses de 511 participants (n=511).

L'objectif est d'identifier et de décrire les différents profils psychologiques existants dans cet échantillon de traileurs.euses.

2.2 Variables mesurées

Ce questionnaire recueille des données liées aux caractéristiques individuelles, sportives et psychologiques. Une partie des variables recueillies par ce questionnaire est récapitulé ci-dessous *(Tableau 1)*.

Tableau récapitulatif des variables mesurées		
Caractéristiques individuelles	*Caractéristiques sportives*	*Santé physique et mentale*
- Sexe	- Temps d'expérience en trail	-Blessure dans l'année écoulée
-Âge	- Entraînement croisé	- Type de Motivation (BRSQ)
- Poids	- Proportion d'entraînement croisé dans l'entraînement	- Force mentale (SMTQ)
- Taille	- Suivi par un coach spécialisé	- Trait d'anxiété pré-compétitive (Sas-2)
- Volume hebdomadaire de travail	- Connaissance point ITRA	- Addiction (EAI)
	- Volume d'entraînement en phase de préparation à une course similaire à celle à laquelle est inscrit le participant à l'UT4M 2024	- Passion (Échelle de la passion)
	- Course faite sur l'Ut4M 2024	

Tableau 1. Présentation des variables recueillies par le questionnaire

Les types de motivation (motivation intégrée, intrinsèque, identifiée, introjectée, externe), la force mentale (consistance, confiance, contrôle), l'anxiété (somatique et cognitive) et la passion (harmonieuse et obsessive) seront les variables utilisées pour identifier les profils psychologiques.

2.3 Identification des profils psychologiques

2.3.1 Analyse factorielle confirmatoire (AFC)

Certaines caractéristiques psychologiques qui ont été mesurées à l'aide du questionnaire, ont pu être regroupées et vérifiées avec une AFC.

L'analyse factorielle confirmatoire sert à vérifier la structure factorielle d'un ensemble de variables observées. Cela permet de tester l'hypothèse de la relation entre la variable observée et ses facteurs latents sous-jacents. L'indice d'ajustement (cfi), lorsqu'il est supérieur à 0.9, témoigne d'un bon ajustement du modèle. La suppression d'items peut être effectuée lors de l'inadéquation de cet ajustement, les items qui contribuent le moins à la variable observée peuvent donc être supprimés pour améliorer l'indice.

Concernant les types de motivation mesurés dans le questionnaire, elles ont été regroupées en motivation autonome et contrôlée puisque ce sont les construits les plus utilisés dans la littérature (Ryan & Deci, 2018[6]). On utilise en effet majoritairement le regroupement sous ces deux formes plutôt que les types de motivation (i.e., identifiée, introjectée...) pris isolément. Les motivations intégrée, intrinsèque et identifiée sont regroupées en motivation autonome et les motivations introjectée et externe en motivation contrôlée. Chaque type de motivation comporte 4 items donc un total de 12 pour la dimension de motivation autonome et un item de la motivation introjectée a été supprimé donc 7 items pour la dimension de la motivation contrôlée (cfi = 0.90). Pour la force mentale, la consistance, le contrôle et la confiance ne sont pas utilisés individuellement mais sont regroupés en une dimension globale, l'objectif n'étant pas de distinguer les trois mais plutôt d'avoir une variable mesurant la force mentale générale. La force mentale comporte donc 13 items avec le contôle qui a 4 items, la consistance 3 items (un a été retiré) et la confiance 6 items (cfi = 0.91). Les types d'anxiétés, étant de nature différente, l'anxiété cognitive est un processus psychologique lié aux pensées et l'anxiété somatique renvoie à des manifestations physiologiques (Martens et al., 1990[4]), ils n'ont pas été regroupés (5 items chacun, cfi = 0.99). Les passions harmonieuses et obsessives étant considérées comme deux construits distincts dans la littératue (i.e., mesurant deux types de

passion différentes) (Vallerand, 2008[9]), ces variables n'ont par conséquent pas été regroupées (6 items chacun, cfi = 0.95). La fidélité des mesures a été validée avec l'omega de McDonald. Ces analyses ont été effectuées à l'aide du package *lavaan* sur R.

2.3.2 Modèle de mélange gaussien fini

Suite à ces premières analyses, un modèle de mélange gaussien a été utilisé pour effectuer le clustering. Uniquement les variables psychologiques, présentées ci-dessus, et préalablement sélectionnées et validées grâce à l'analyse factorielle ont été utilisées.

Le modèle de mélange consiste à modéliser un ensemble de données comme étant composé de plusieurs clusters qui suivent une distribution gaussienne. Soit $x = \{x_1,x_2,...,x_n\}$ un échantillon de n observations indépendantes et identiquement distribuées. Chaque observation est associée à une fonction de densité de probabilité à l'aide d'un modèle de mélange fini

$$f\left(x_{i;}\Psi\right) = \sum_{k=1}^{G} \pi_k f_k(x_i;\theta_k) \quad (1)$$

avec G le nombre de clusters, $\Psi = \{\pi_1,...,\pi_{G-1},\theta_1,...,\theta_G\}$ les paramètres du modèle, $f_k(x_i;\theta_k)$ la k-ième densité composante pour x_i avec θ_k le vecteur de paramètres et $(\pi_1,...,\pi_{G-1})$ les poids de mélange. Dans le cas du mélange gaussien, $f_k\left(x_i;\theta_k\right) \sim N(\mu_k,\Sigma_k)$ où μ_k est le centre du cluster et Σ_k la matrice de covariance qui donne entre autre le volume ou la forme des clusters (Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E., 2016[8]). Les paramètres Ψ sont estimés en maximisant la log-vraisemblance à l'aide de l'algorithme d'Expectation-Maximization (EM), ce qui permet d'obtenir une probabilité d'appartenance des observations aux clusters. Le clustering a été effectué à l'aide du package *Mclust* sur R.

Le modèle de mélange a été testé avec le nombre de clusters G allant de 2 et 10. Le critère BIC (Bayesian Information Criterion) a permis de sélectionner le meilleur modèle, avec 4 groupes.

2.4 Description des profils psychologiques

Afin d'avoir une première visualisation des caractéristiques des profils obtenus, les scores moyens des clusters sur les différentes variables psychologiques sont représentés *(Figure 1)*.

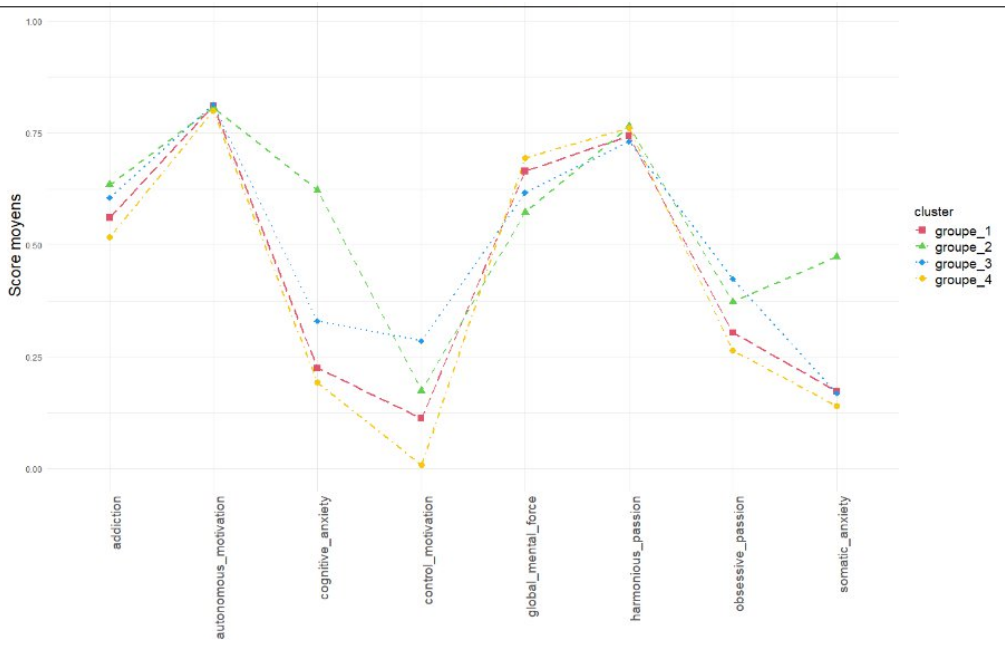


Figure 1. Scores moyens des clusters sur les différentes variables psychologiques étudiées

L'analyse par clustering fait émerger des profils psychologiques distincts. Certaines variables (i.e., anxiété cognitive, anxiété somatique, passion excessive) semblent jouer un rôle important dans la détermination des différents clusters. Nous présenterons l'interprétation de ces résultats.

3 Perspectives

Les prochains objectifs seront d'examiner si ces profils peuvent expliquer la blessure et le risque d'addiction au trail ainsi que d'évaluer si les pratiquants d'ultra-trail (i.e., distance de course égale ou supérieure à 100 km) se distinguent des pratiquants de distance plus courte d'un point de vue du profil psychologique notamment à l'aide de tests statistiques.

Bibliographie

(1) Béland, S., Cousineau, D., & Loye, N. (2018). Utiliser le coefficient omega de McDonald à la place de l'alpha de Cronbach. *McGill Journal Of Education / Revue des Sciences de L Éducation de McGill, 52(3)*, 791-804. https://doi.org/10.7202/1050915ar

(2) Bonneville-Roussy, A., Fenouillet, F., & Morvan, Y. (2022). *Introduction aux analyses par équations structurelles.* Dunod.

(3) *ITRA - International Trail Running Association - official world ranking of trail runners recognized by World Athletics. (s. d.).* https://itra.run/

(4) Martens, R., Vealey, R. S., & Burton, D. (1990). Competitive anxiety in sport. Champaign, IL: Human Kinetics.

(5) Plard, M., & Martineau, A. (2021). Le trail-running à l'épreuve des sciences, une revue de littérature interdisciplinaire. Panorama des sciences humaines, sociales et médicales. https://hal.science/hal-03196251

(6) Ryan, R. M., & Deci, E. L. (2017). Self-determination theory: Basic psychological needs in motivation, development, and wellness. The Guilford Press. https://doi.org/10.1521/978.14625/28806

(7) Roebuck, G. S., Fitzgerald, P. B., Urquhart, D. M., Ng, S.-K., Cicuttini, F. M., & Fitzgibbon, B. M. (2018). The psychology of ultra-marathon runners : A systematic review. Psychology of Sport and Exercise, 37, 43-58. https://doi.org/10.1016/j.psychsport.2018.04.004

(8) Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016b, août 1). mclust 5 : Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. https://pmc.ncbi.nlm.nih.gov/articles/PMC5096736/#S4

(9)Vallerand, Robert & Blanchard, Céline & Mageau, Geneviève & Koestner, Richard & Ratelle, Catherine & Léonard, Maude & Gagné, Marylène & Marsolais, Josée. (2003). Les Passions De L'ame: On Obsessive and Harmonious Passion. Journal of personality and social psychology. 85. 756-67. 10.1037/0022-3514.85.4.756.

ÉTUDE DE MODÈLES DE DIFFUSION POUR LA MODIFICATION D'IMAGES DE DANSE SPORTIVE

Lucas Bondietti [1], Ismaël Viennot [1], Juliette Chevallier [1,2] & Emmanuelle Claeys [3,4]

[1] *INSA Toulouse, Université de Toulouse, France*
[2] *Institut de Mathématiques de Toulouse, UMR 5219*
[3] *Université Paul Sabatier, Université de Toulouse, France*
[4] *Institut de Recherche en Informatique de Toulouse, UMR 5505*

Résumé. La génération d'images et de vidéos à travers des modèles génératifs et de diffusions est un domaine très actif de la recherche, notamment à travers la popularisation de modèles tels que Chat-GPT ou LAMA. L'augmentation des ressources de calcul a permis ces dernières années le développement de modèles capables de générer de nouvelles images et vidéos à partir de sources originales, ou encore de modifier ces dernières en suivant des contraintes fournies sous forme de prompts. Pourtant, ces modèles ont encore des difficultés à préserver l'anatomie humaine lorsqu'on les applique à des images présentant des corps humains effectuant des mouvements complexes, tels que ceux présents en gymnastique ou en danse. Dans cet article, nous proposons un retour d'expérience sur l'utilisation des dernières méthodes de référence pour de la modification d'images de danse artistique, ainsi qu'un nouveau cadre d'étude basé sur du *Poisson Blending* permettant un meilleur respect de la morphologie des danseurs.

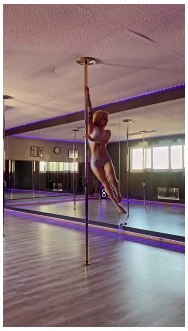
Mots-clés. Machine learning pour l'analyse sportive, modèles de diffusion, modèles génératifs, danse sportive.

Abstract. Generating images and videos using generative and diffusion models is a very active area of research, notably through popularizing models such as Chat-GPT or LAMA. In recent years, increased computational resources have enabled the development of models to generate new images and videos from original sources, or modify them according to constraints provided by prompts. However, these models still have difficulty in preserving human anatomy when applied to images displaying human bodies performing complex movements, such as those found in gymnastics or dance. In this article, we provide experiments relying on the latest reference methods for image modification in the context of artistic dance, as well as a new framework based on *Poisson Blending* allowing better respect of dancers' morphology.

Keywords. Machine learning for sports analysis, diffusion models, generative models.

1 Introduction

Dans le contexte de la danse artistique et sportive, les athlètes sont très souvent amenés à faire la promotion de leur activité via la diffusion de courtes vidéos. Ces vidéos sont généralement capturées en intérieur, dans des studios de danse et sur un dispositif photographique de faible qualité, typiquement un téléphone portable. Aussi, leur qualité artistique reste limitée.



(a) Julie Cuirans

(b) Didier Benac

Figure 1: Deux athlètes professionnels photographiés dans leur studio respectif.

L'objectif de ce travail est de permettre la modification de ces vidéos, notamment de leur fond, à l'aide d'un *prompt* de commandes textuelles, et ceci afin d'améliorer l'esthétisme de ces productions. Nous nous concentrons ici sur la pole dance ; mais, les techniques développées s'appliquent directement à toutes disciplines présentant des caractéristiques similaires : gymnastique, danse plus généralement, *etc.*

On trouve dans la littérature beaucoup de méthodes open-source permettant de modifier une image ou bien une vidéo à partir d'un *prompt* [9]. Nos premiers essais ont montré que ces modèles sont performants sur des positions simples ou des animaux ; mais, lorsqu'ils sont utilisés sur des mouvements assimilés à de la gymnastique, ils conduisent à des artefacts et dénaturent complètement le corps humain : on parle de *body horror*. Ce résultat était attendu. En effet, les modèles génératifs ont encore beaucoup de difficultés à comprendre la biophysique basique, et cela reste une des limitations à leur usage [10].

Dans cet article, à partir de deux photos de danseurs (*cf.* Fig. 1), nous proposons un retour d'expérience ainsi qu'un nouveau cadre permettant l'édition du fond des photos, tout en préservant autant que possible le corps des deux athlètes. La section 2 propose un bref état de l'art sur les modèles de diffusion, la section 3 détaille notre nouveau cadre et évalue sa performance sur les deux images de référence susmentionnées.

2 État de l'art

Les modèles de diffusion [4, 17, 18] sont une catégorie de modèles génératifs basés sur des réseaux de neurones convolutionels (CNNs, [6]). Étant donnée une image originale, ces modèles fonctionnent en deux phases :

- ***Forward diffusion*** : L'image est progressivement bruitée par l'ajout progressif et markovien d'un bruit gaussien ;
- ***Reverse diffusion*** : À partir de l'image complètement bruitée, le modèle la débruite progressivement pour reconstruire une image réaliste, c'est-à-dire labellisée comme conforme par un classifieur.

Nous décrivons ci-après plus en détail le fonctionnement de ces réseaux.

Entraînement. Pour chacune des images composant le jeu d'entraînement, un bruit gaussien leur est ajouté, en plusieurs étapes $t \in [\![1, T]\!]$. On note x_0 l'image avant modifications, x_t l'image obtenue à la t-ième étape du processus. En particulier, x_T est complètement bruitée et ininterprétable. Plus précisément, le processus de *forward diffusion* est donné par l'équation

$$x_t \;=\; \sqrt{1-\beta_t}\, x_{t-1} \;+\; \sqrt{\beta_t}\, \varepsilon_t \,,$$

où β_t contrôle le niveau de bruit à chaque étape, et $\varepsilon_t \sim \mathcal{N}(0, I)$. Ainsi, par récurrence, on peut définir x_t à partir de x_0 :

$$x_t \;=\; \left(\prod_{s=1}^{t} \sqrt{1-\beta_s} \right) x_0 \;+\; \sum_{k=1}^{t} \left(\sqrt{\beta_k} \prod_{s=k+1}^{t} \sqrt{1-\beta_s} \right) \varepsilon_k \,. \tag{1}$$

Considérons un réseau de neurones paramétrisé par θ et visant à approcher ce processus. Historiquement [17], ces réseaux étaient entraînés par inférence variationelle via une perte ELBO, ou *evidence lower bound*, portant sur une approximation par le réseau de la loi de x_0. Des travaux plus récents [4] ont permis d'obtenir de meilleurs résultats empiriques en contrôlant l'erreur quadratique moyenne, ou *mean squared error* (MSE), entre le bruit réel associé à l'image x_t et celui prédit par le réseau noté $\varepsilon_\theta(x_t, t)$. Remarquons qu'xen notant $\bar{\beta} := \left(\prod_{s=1}^{t} \sqrt{1-\beta_s}\right)$ et $\sigma_t := \sum_{k=1}^{t} \left(\sqrt{\beta_k} \prod_{s=k+1}^{t} \sqrt{1-\beta_s}\right)$, par propriété des vecteurs gaussiens, l'équation (1) se réécrit $x_t = \bar{\beta}\, x_0 + \sigma_t\, z$, où z désigne un bruit gaussien inconnu, et on remarque bien une symétrie entre la compréhension de x_0 et celle de z. Ceci conduit à la perte dite simple

$$\mathcal{L}\,(\theta) \;:=\; \sum_{t=1}^{T} \mathbb{E}_{x_0 \sim q,\, z \sim \mathcal{N}(0,I)} \left[\left\| \varepsilon_\theta(x_t, t) - z \right\|^2 \right] .$$

Inférence. Après entraînement, un modèle de diffusion est à même de générer de nouvelles observations réalistes, ou de modifier des images existantes, ceci via un usage judicieux du processus de *reverse diffusion* précédemment décris.

La génération pure s'effectue à partir d'un bruit blanc gaussien x_T, débruité étape par étape. Autrement dit, au cours de T étapes, le modèle prédit et supprime incrémentiellement le bruit $\varepsilon_\theta(x_t, t)$ de l'image x_t :

$$x_{t-1} \;=\; \frac{1}{\sqrt{1-\beta_t}} \big(x_t - \beta_t\, \varepsilon_\theta(x_t, t) \big) + \sqrt{\beta_t}\, \zeta \,, \tag{2}$$

où ζ est un bruit gaussien qui assure la diversité dans les modifications.

Par opposion, pour la modification d'image, l'état initial bruité x_T est conditionné par l'image originale ayant été labellisée automatiquement. Les modifications apportées à l'image peuvent également être conditionnées par une base d'entrées externes, telles que des descriptions textuelles ou *prompts*. Cela permet au modèle de conserver des éléments de l'image originale, tout en appliquant les modifications souhaitées. Concrètement, en notant c l'entrée de conditionnement, le processus de *reverse diffusion* décrit Eq. (2) est inchangé, sauf à tenir compte de c dans la définition du bruit prédit $\varepsilon_\theta(x_t, t, c)$.

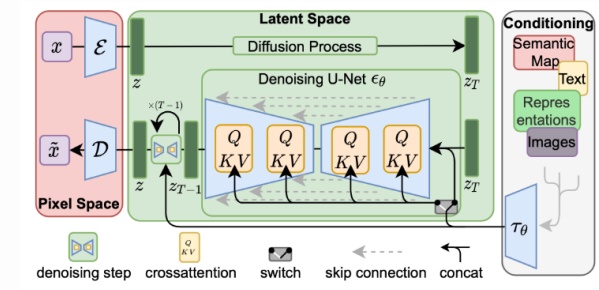


Figure 2: *Latent diffusion models* ou LDM – Image tirée de [15].

À droite, les contraintes de conditionnement sont intégrées à l'image originale, conduisant à une représentation latente combinée z_t hautement sémantique. La diffusion porte alors sur ces représentations latentes.

Dans notre contexte, nous souhaitons éditer le fond d'une photo de danse, tout en préservant le corps des athlètes, et en contrôlant la manière dont est modifiée le reste de l'image. Pour ce faire, nous allons recourir à deux types de conditionnement :

- **Descriptions textuelles** ou *prompts* : Afin d'être interprétées, les descriptions textuelles sont préalablement transformées en vecteurs numériques multidimensionnels par *embedding*. Ceci conduit à une représentation combinée de l'image, qui peut ensuite être utilisée pour préserver des contraintes dans chaque étape du modèle de diffusion.
- **Caractéristiques explicites** ou *feature conditioning* : Afin de fournir des contraintes localisées, on peut recourir à l'usage de masques de segmentation, de cartes de profondeur ou de bords. Le modèle est ainsi contraint de respecter des informations structurelles ou spatiales encapsulées dans ces cartes de caractéristiques ou *features maps*, garantissant l'alignement de l'image finale sur les conditions spécifiées.

Ces *features* de conditionnement c sont combinées à la représentation latente de l'image bruitée x_t par le biais de mécanismes d'attention [7, 19], modifiant l'estimateur de bruit ε_θ comme suit :

$$\varepsilon_\theta(x_t, t, c) \;=\; \texttt{CrossAttention}\left(\varepsilon_\theta(x_t, t)\,,\, c\right).$$

Nous appelons ces modèles mêlant mécanismes d'attention et processus de diffusion modèles de diffusion latente ou *latent diffusion models* (LDM, [15]). Le schéma reproduit Fig. 2 et issus de l'article séminale [15] décrit leur architecture. L'idée de ces réseaux est de forcer le respect des contraintes spatiales via l'usage d'un réseau de type U-Net [16]. En particulier, l'attention vise ici à s'assurer de la présence de certaines "clés" (*keys*) K dans la représentation latente successivement modifiée par l'U-Net. Les clés à maintenir sont définies par un ensemble de "requêtes" (*queries*) Q et sont chacunes associées à une valeur V. Concrètement, $\texttt{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \cdot V$, pour une représentation latente de dimension d et des matrices Q, K, V à estimer.

3 Contribution

3.1 Retour d'expérience

Nous présentons dans cette section les résultats que nous avons obtenus sur trois modèles de diffusion latente ou LDMs, plébiscités pour leurs performances :

- *Stable Diffusion XL* (SDXL, [12]) incorpore des couches supplémentaires au LDM et améliore l'U-Net sous-jacent. Cette architecture, plus profonde, permet un traitement plus détaillé et multi-échelle des informations par SDXL, conduisant à la génération d'images de meilleure qualité, avec un niveau de détails plus fins et une meilleure gestion des textures complexes par rapport à un LDM standard.
- *Image-to-Image* [20] permet l'ajout de paramètres optionnels tels que le nombre d'étapes d'inférence ou la force de la transformation. Plus précisément, le paramètre "force" contrôle la déviation du modèle par rapport à l'image d'entrée : une valeur faible maintient la sortie proche de l'original, tandis qu'une valeur plus élevée autorise des changements plus importants.
- *Depth-to-Image* [14] estime la distance entre l'appareil photo et les objets présents dans l'image en analysant les pixels, conduisant à la création d'une carte de profondeurs faisant office de conditionnement pour le mécanisme de diffusion. Cela permet la conservation de la structure spatiale des images lors de leur modification.

Les meilleurs résultats obtenus par ces modèles sont fournis en annexe, Figs. 4, 5, et 6. À noter qu'aucun des modèles précédemment listés n'a produit de résultats concluants. Nous intuitons que cela provient de leur entraînement : ces modèles ne sont pas entraînés sur un grand ensemble d'images de danseurs ou de gymnastes, ce qui limite leur compréhension de l'anatomie des athlètes en mouvement. Dans des cas plus simples, tels que l'édition d'images d'animaux ou de paysages, les résultats obtenus étaient alors tout à fait satisfaisants.

Dans la section suivante, nous proposons un nouveau *framework* couplant l'utilisation d'un masque pour préserver le corps des athlètes, la segmentation de ce dernier, la complétion de l'image évidée par *inpainting*, et enfin la fusion du masque et de l'image complétée, tout en préservant la cohérence de l'image finale.

3.2 Nouveau cadre de travail

La première étape consiste en la localisation et la segmentation automatique du danseur de l'image à l'aide de *Segment Anything* (SAM, [5]). Étant donnée une image en couleurs de dimension $h \times w$, où h et w représentent respectivement la hauteur et la largeur, l'image est divisée en n patchs [8] de dimensions $p \times p$ tels que $n = \frac{hw}{p^2}$. Les patchs sont redimensionnés et projetés linéairement par une matrice de poids W_p, de sorte à définir le vecteur de caractéristiques x_i associé au i-ième patch P_i par :

$$x_i \;=\; W_p \cdot \texttt{Flatten}(P_i) + E_{pos,i}\,,$$

(a) Masques de ségmentation obtenu par SAM (b) Poisson Blending

Figure 3: Segmentation et remplissage des deux images d'origine par *Poisson Blending*.

où $E_{pos,i}$ désigne le codage positionnel et correspond à l'information spatiale. Ces vecteurs de caractéristiques sont alors traités comme des séquences d'entrée par un transformeur afin de produire des représentations des caractéristiques de l'image tenant compte du contexte. Les interactions entre les patchs sont modélisées à l'aide de la fonction `CrossAttention` décrite section 2. Le résultat de cette segmentation pour nos images de références est donnée Fig. 3a.

Une fois la segmentation réalisée, l'image évidée de l'athlète est complétée par *inpainting* [1, 13] ; puis, l'image ainsi complétée est modifiée via *Stable Diffusion XL* (*cf.* Section 3.1) conditionné par le masque de segmentation. Ce conditionnement permet de s'assurer de la cohérence et de la conservation de la structure globale de l'image, les parties non masquées agissant alors comme des contraintes. À ce stade, l'image modifiée consiste en une représentation alternée, mais cohérente, du fond de l'image (ici, un studio de danse). En particulier, l'athlète n'est plus présent sur l'image.

Afin de réintégrer les danseurs de manière cohérente dans les images éditées, nous procédons par *Poisson Blending* [11]. L'idée du *Poisson Blending* est de copier des zones d'une image source S (ici, l'athlète) dans une image cible T (ici, le studio modifié par SDXL) tout en imposant des contraintes de recollement et de régularité. Plus précisément, le résultat du *Poisson Blending* est l'image x obtenue via l'équation suivante :

$$\min_x \int_\Omega ||\nabla x - \nabla S||^2 \quad \text{tel que} \quad x_{\mathcal{D}\setminus\Omega} = T\,,$$

où Ω est un masque binaire représentant la zone de S à copier dans T et $\mathcal{D}$ est le domaine des images. Intuitivement, la régularité du recollement provient du fait que l'on copie les gradients spatiaux ∇S de l'image source dans T, et non les valeurs des pixels composants S. Cette équation est résolue à l'aide de solveurs itératifs classiques, par exemple par méthode du gradient conjugué [3]. Les résultats obtenus par cette méthode sont présentés Fig. 3b.

4 Conclusion

Nos expériences ont mis en lumière la difficulté que représente la modification d'une image, tout en conservant l'anatomie de corps en mouvements techniques, tels que ceux rencontrés en danse. En particulier, même les algorithmes à l'état de l'art pour la modification d'images ne parviennent pas à réaliser de telles modifications. Néanmoins, en combinant plusieurs méthodes différentes, nous avons pu obtenir des résultats satisfaisants.

Les perspectives de ce travail sont multiples. On pourra par exemple s'intéresser au passage de la photo à la vidéo, ce qui soulèvera de nouveaux verrous méthodologiques : En effet, les modifications d'un plan à l'autre de la vidéo devront alors être également cohérentes entre elles. Une autre piste de travail envisagée consiste en l'intégration des contraintes biophysiques directement dans le modèle de diffusion, probablement via du conditionnement.

References

[1] Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting

[2] Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023)

[3] Hestenes, M.R., Stiefel, E., et al.: Methods of conjugate gradients for solving linear systems, vol. 49. NBS Washington, DC (1952)

[4] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[5] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)

[6] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)

[7] Lin, H., Cheng, X., Wu, X., Shen, D.: Cat: Cross attention in vision transformer. In: IEEE international conference on multimedia and expo (ICME). pp. 1–6 (2022)

[8] Nguyen, D.K., Assran, M., Jain, U., Oswald, M.R., Snoek, C.G., Chen, X.: An image is worth more than 16x16 patches: Exploring transformers on individual pixels. Preprint arXiv:2406.09415 (2024)

[9] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. Preprint arXiv:2112.10741 (2021)

[10] O'Meara, J., Murphy, C.: Aberrant ai creations: co-creating surrealist body horror using the dall-e mini text-to-image generator. Convergence **29**(4), 1070–1096 (2023)

[11] Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 577–582 (2023)

[12] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. Preprint arXiv:2307.01952 (2023)

[13] Qin, Z., Zeng, Q., Zong, Y., Xu, F.: Image inpainting based on deep learning: A review. Displays **69**, 102028 (2021)

[14] Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)

[15] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

[16] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)

[17] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)

[18] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. Preprint arXiv:2011.13456 (2020)

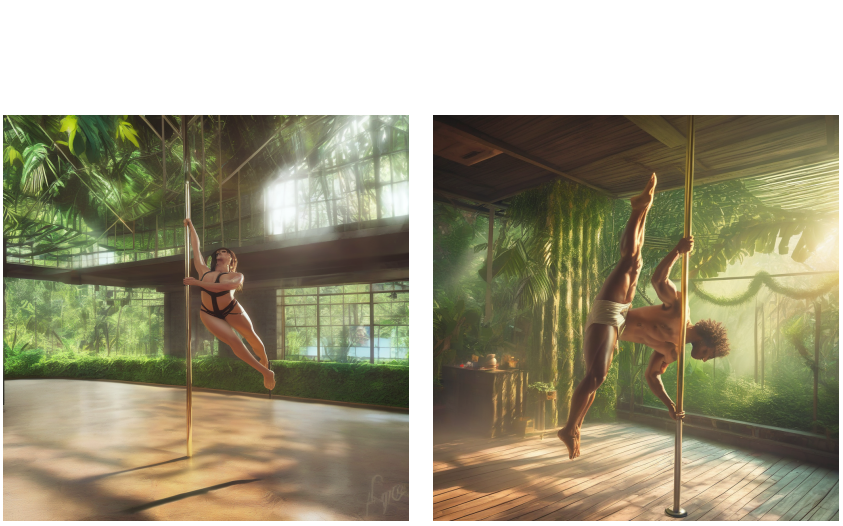
[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems (2017)

[20] Yu, X., Tian, J., Hu, Z.: An analysis for image-to-image translation and style transfer. Preprint arXiv:2408.06000 (2024)

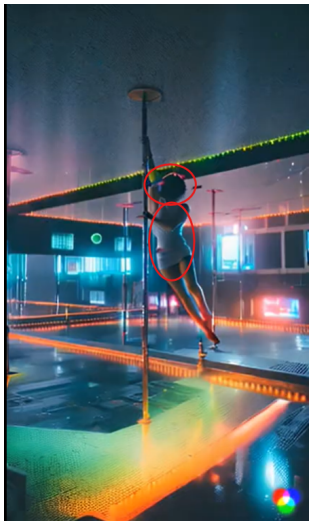
Annexe

Figure 4: Images modifiées à l'aide d'*Image-to-Image* [20].

L'utilisation d'*Image-to-Image* sur des corps d'athlètes en mouvement conduit à des *body horror*.



(a) Images modifiées par *Depth-to-Image*



(b) Image modifiée par *GEN2*

Figure 5: Images modifiées par *Depth-to-Image* [14] et *GEN-2* [2].

Ces deux réseaux, bien qu'à l'état de l'art pour la modification d'images, altèrent la morphologie des athlètes: *Cf.* le visage et le buste Fig. 5a-Gauche, la musculature Fig. 5a-Droite, les cercles rouges Fig. 5b.

(a) `A woman doing pole dance in the jungle`

(b) `A man doing pole dance in the jungle`

Figure 6: Images générées par SDXL [12], à l'aide d'un texte simple.

La légende de chaque image reporte le *prompt* utilisé pour la génération. Ces exemples tendent à montrer que SDXL est capable de traiter ce type de données.

Rugby Union as electric consumption : Short-term performance prediction in team sports using tracking data

Arnaud Odet [1,2], Sébastien Déjean [1] & Cristian Pasquaretta [2]

[1] *Institut de Mathématiques de Toulouse, UMR 5219 Université de Toulouse et CNRS,* {*arnaud.odet, sebastien.dejean*}*@math.univ-toulouse.fr*
[2] *Centre de Recherches sur la Cognition Animale (CRCA), Centre de Biologie Integrative (CBI), CNRS, Université de Toulouse, cristian.pasquaretta@univ-tlse3.fr*

Résumé. Cette recherche porte sur la prédiction à court terme de la performance en Rugby à XV en utilisant des données de *tracking* des joueurs. Les caractéristiques uniques du rugby en tant que sport de conquête avec des règles strictes de hors-jeu font de la position du ballon un indicateur crucial de performance. Cependant, les données de suivi du ballon sont souvent indisponibles.

Nous avons d'abord évalué différentes méthodes d'estimation de la position du ballon en fonction de la position des joueurs, puis, sur cette base, nous cherchons à prédire l'évolution de la position du ballon, utilisée comme indicateur de performance, à partir des données de *tracking*. Nous avons formulé cette tâche comme un problème de prédiction de séries temporelles où les positions des joueurs (multivariées) prédisent la position du ballon (univariée).

Mots-clés. Données de *tracking*, Series temporelles, Prediction de trajectoire multi-agents, Sport, Apprentissage automatique.

Abstract. This research addresses short-term performance prediction in Rugby Union using player tracking data. Rugby's unique characteristics as a conquest sport with strict offside rules make the ball position a crucial performance indicator. However, ball tracking data is often unavailable.

We first evaluated different methods to estimate ball position as a function of player positions, and then, building on this foundation, we aim to predict ball position evolution, used here as performance indicator, using player tracking data. We formulated this challenge as a time series prediction problem where player positions (multivariate) predict ball position (univariate).

Keywords. Tracking Data, Time Series, Multi-Agent Trajectory Prediction, Sport, Machine Learning.

1 Introduction

The prediction of performance in sports has become increasingly important with the advancement of player tracking technologies and analytics. Performance can be defined and analysed at different scales : over a complete season, over a game, or over an action. This paper focuses on **short-term performance prediction** (i.e. over an action), in team sports, specifically in Rugby Union, where the dynamic nature of play presents unique analytical opportunities and challenges.

This objective can be formulated as a Time Series prediction task, where multivariate (tracking data) predicts univariate (performance). This prediction task is particularly challenging due to the complex nature of team sports data, which involves multiple agents (players) working collaboratively with role-based relationships towards a common objective.

Rugby Union, as a conquest sport, is characterized by strict rules governing player positioning, particularly the offside rule, which makes the ball position a crucial indicator of territorial advantage and play development. This characteristic suggests that tracking the ball's position can serve as an effective proxy for performance assessment and prediction. However, ball tracking data is often unavailable or incomplete in many professional sports settings, including Rugby Union, creating a significant challenge for performance analysis.

To address this limitation, our study first conducted a systematic evaluation of different ball position estimation methods. We compared various estimation approaches against a subset of manually annotated ground truth data. This evaluation enabled us to identify and validate the most accurate method for ball position estimation, which was subsequently applied to all matches where direct ball tracking was unavailable.

We then used a comprehensive dataset comprising 22 professional Rugby Union games captured at 10Hz frequency. The high-frequency sampling of our dataset captures the rapid movements and tactical developments characteristic of Rugby Union, adding another layer of complexity to the prediction task.

The implications of this work extend beyond immediate performance prediction in Rugby Union. Our methodology for ball position estimation addresses a common challenge in sports analytics, while our prediction frameworks contribute to the understanding of team coordination and tactical decision-making. These findings have potential applications in both real-time game analysis and post-match performance evaluation, providing valuable tools for coaches, analysts, and sports scientists. Moreover, our approach could be adapted to other team sports with similar structural characteristics, contributing to the broader field of sports analytics and performance prediction.

2 Related work

The use of tracking data for sports performance prediction has advanced significantly with the development of player-tracking technologies.

Tracking Data for Sports Performance Prediction: Initial work with tracking data involved quantifying physical performance and time-motion analysis (Clemente et al., 2013), with a subsequent shift toward using this data for tactical analysis and understanding collective movement patterns (Low et al., 2020; Perše et al., 2009).

A key area of research involves creating performance indicators derived from spatial and temporal aggregations of tracking data. These often include metrics of space control, team centroid, and other features based on domain expertise (Clemente et al., 2013; Low et al., 2020).

Machine learning is increasingly used to automate the identification of tactical patterns, using trajectory data and reducing reliance on manual annotation (Anzer et al., 2022). This includes detecting key events and summarizing team and player actions automatically (Tian et al., 2019). More recent studies focus on rating specific actions based on their impact on the game, estimating the probability of a successful pass or shot (Dick et al., 2021).

Time Series Forecasting: Time series analysis involves studying collections of observations ordered by time, with applications across various fields including energy, economics, and meteorology. The main challenge is to understand the temporal dynamics in Time Series, which includes decomposition, seasonality, long-term dependencies, etc. It has been a very dynamic, long-standing field of study (Klein, 1997), with traditional statistical methods (ARMA and descendants, see Box et al., 1978) and advances in Deep Learning with RNN (Rumelhart et al., 1986), LSTM (Hochreiter & Schmidhuber, 1997), Transformer (Vaswani et al., 2017) and more recently Mamba (Gu & Dao, 2023). Transformer models have been adapted for time series analysis, with a focus on their ability to capture long-range dependencies, with a wide focus on the attention mechanism (see review by Wang et al., 2024). PatchTST (Nie et al., 2023) improves long-term forecasting accuracy by segmenting time series into subseries-level patches. TimesNet (Wu et al., 2023) is a task-general foundation model for time series analysis that models temporal variations in a 2D space. Non-stationary Transformers (Liu et al., 2022) uses methods like De-stationary Attention to handle non-stationarity in time series, outperforming stationarization techniques. Zeng et al., 2022 questioned the necessity of using Transformers for long-term forecasting, suggesting that simpler linear models can also achieve competitive results.

3 Ball position estimation

As Rugby Union is a conquest sport and imposes a strict offside rule, the position of the ball on the x axis of the field is a good indicator of desirable outcome. However, despite recent technological advances in ball tracking technology, these technologies remain under-explored notably due to prohibitive costs.

We therefore used a set of 3 games with manually annotated ball location, downsampling our tracking frenquency to 2Hz, leaving us with 3426 frames that were split in train set (70%), validation set (15%) and test set (15%). The ball position was obtained by labelling the player holding the ball in offense, and is therefore only available when the team of interest is in possession of the ball. We compared different types of models among the most commonly used (linear regression, ensemble methods, support vector machine, k-NN, and ANN, as well as 3 "naive" methods). All these methods were applied twice : (i) on player identity, i.e. position on the field, and (ii) player "pseudo-rank" where players are ranked by proximity to the opponent try line (to best leverage the offside rule). To optimise the performance of these algorithms, we conducted a random search for the hyperparameters over standard predefined ranges.

Surprisingly, the method that produces the lower RMSE is the naive method using the position on the x axis of the player having the lowest average distance to the ball in the train set as an estimator of the position of the ball. This method is equivalent to considering the position over the x axis of player number 9 (scrum half). The identity of the identified player is intuitive, as the scrum half's role is to be always close to the ball, being the link between the pack (numbers 1 to 8) and the other players. On the other hand, the fact that no algorithm outperforms this naive method is noteworthy : this method displays an RMSE at least 18% lower than any other model we evaluated.

4 Short-term performance prediction in Rugby Union

In this section, we used a dataset containing 22 PROD2 games of a given team, with a frequency of 10 Hz. We only considered offensive situations as our ball estimation method was learned from situations where the team analysed is in possession of the ball.

Let $X_{(j,t)}$ be the feature vector of player $j \in \{1, ..., 15\}$ at time t. $X_{(j,t)}$ is 4-dimensional, as it contains the position (x and y coordinates) and the velocity (v_x and v_y) of player j. We note $X_t = \{X_{(1,t)}, ..., X_{(15,t)}\}$ the feature vector at time t, s the sequence length and p the prediction length. Let B_t be the position of the ball over the x axis (estimated or not) at time t, and $S_t = B_{t+s+p} - B_{t+s}$ be the performance variable associated with the sequence $\{X_t, ..., X_{t+s}\}$. It is obtained by substracting the x coordinate of the ball at the start of the prediction period to the the x coordinate of the ball at the end of the prediction period, materializing the progress made during the prediction period.

Our objective can be formulated as **predicting** S_t **given** $\{X_t, ..., X_{t+s}\}$. Alternatively, it can be reformulated as : Predicting $\{B_{t+s+1}, ..., B_{t+s+p}\}$ given $\{X_t, ..., X_{t+s}\}$ (multivariate

to univariate Time Series prediction task), where methods are compared using the Final Displacement Error (FDE) metric, as the FDE only considers the last point of a trajectory.

Formally, the FDE is defined as $FDE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(\hat{B}_{t_i+s+p} - B_{t_i+s+p})^2}$ where $\hat{B}$ is the predicted position of the ball and t_i indicates the start of sequence $i, i \in \{1, ...n\}$.

We arbitrary set $s = p = 32$, and removed all sequences where no game action occurs, as well as all frames with missing or aberrant value. We extracted sequences with an overlap of 4, generating 28 116 sequences that were split between training (70%), validation (15%) and test (15%) sets.

We compared the following methods : 2 naive baselines (the first considering the ball does not move during the prediction window, the second assuming it moves at a constant velocity, being the last observed velocity), 2 vanilla models (LSTM and Transformer, see Hochreiter and Schmidhuber, 1997; Vaswani et al., 2017) and several state-of-the-art Time Series models, highlighted by Wang et al., 2024 : iTransformer (Liu et al., 2023), TimesNet (Wu et al., 2023), PatchTST (Nie et al., 2023), Non-Stationary Transformer (Liu et al., 2022).

5 Discussion and conclusion

This study presents several novel contributions to the field of sports analytics. First, we proposed and validated a method evaluating performance by estimating ball position from tracking data in Rugby Union addressing a common limitation in professional sports analysis. Second, we leveraged existing work in Time Series analysis to demonstrate the feasibility of applying these models to predict short-term performance in Rugby Union.

Our work opens up several promising research directions. First, as we used a deterministic function from tracking data to estimate a performance variable, we could aim at predicting future trajectories of player and infer performance from predicted trajectories. This formulation relates our research with multi-agent trajectory prediction, a rapidly evolving field in machine learning, which could make us benefit from deep learning architectures specifically designed for spatiotemporal data.

On the other hand, the growing availability of ball tracking technologies in professional sports presents exciting opportunities for future research. As more teams and competitions adopt advanced tracking systems, the accuracy of ball position, used here as a proxy for performance, will improve.

Additionally, applying explainability techniques, or generating counterfactuals could lead to valuable insights for coaching staffs.

References

Anzer, G., Bauer, P., Brefeld, U., & Fassmeyer, D. (2022). Detection of tactical patterns using semi-supervised graph neural networks.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (1978). Time series analysis: Forecasting and control. *The Statistician*, *27*.

Clemente, M. F., Couceiro, S. M., Martins, F. M. L., Mendes, R., & Figueiredo, A. J. (2013). Measuring Collective Behaviour in Football Teams: Inspecting the impact of each half of the match on ball possession. *International Journal of Performance Analysis in Sport*, *13*(3).

Dick, U., Tavakol, M., & Brefeld, U. (2021). Rating Player Actions in Soccer. *Frontiers in Sports and Active Living*, *3*.

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *ArXiv*, *abs/2312.00752*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*.

Klein, J. L. (1997). Statistical visions in time: A history of time series analysis, 1662-1938.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2023). Itransformer: Inverted transformers are effective for time series forecasting. *ArXiv*, *abs/2310.06625*.

Liu, Y., Wu, H., Wang, J., & Long, M. (2022). Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting.

Low, B., Coutinho, D., Gonçalves, B., Rein, R., Memmert, D., & Sampaio, J. (2020). A Systematic Review of Collective Tactical Behaviours in Football Using Positional Data. *Sports Medicine*, *50*(2).

Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023, March). A Time Series is Worth 64 Words: Long-term Forecasting with Transformers.

Perše, M., Kristan, M., Kovačič, S., Vučkovič, G., & Perš, J. (2009). A trajectory-based analysis of coordinated team activity in a basketball game. *Computer Vision and Image Understanding*, *113*(5).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*.

Tian, C., De Silva, V., Caine, M., & Swanson, S. (2019). Use of Machine Learning to Automate the Identification of Basketball Strategies Using Whole Team Player Tracking Data. *Applied Sciences*, *10*(1).

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Neural Information Processing Systems*.

Wang, Y., Wu, H., Dong, J., Liu, Y., Long, M., & Wang, J. (2024, July). Deep Time Series Models: A Comprehensive Survey and Benchmark.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., & Long, M. (2023). TIMESNET: TEMPORAL 2D-VARIATION MODELING FOR GENERAL TIME SERIES ANALYSIS.

Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2022, August). Are Transformers Effective for Time Series Forecasting?

Statistique et Sport 2

Appréhender certains facteurs qui déterminent la valeur marchande des joueurs de football

Yvenn Amara-Ouali [1,2]

[1] *University Paris-Saclay, LMO, France, yvenn.amara-ouali@universite-paris-saclay.fr*
[2] *EDF R&D, OSIRIS, Palaiseau, France*

Résumé. Le marché mondial des transferts de joueurs de football, un écosystème de plusieurs milliards de dollars, exige une évaluation précise des joueurs pour naviguer dans ses complexités économiques. Les méthodes traditionnelles, qui s'appuient sur des négociations heuristiques et des mesures de performance de base, ne parviennent souvent pas à saisir les interactions non linéaires entre des facteurs tels que les performances sur le terrain, les tendances démographiques, les contraintes contractuelles et les dynamiques macroéconomiques. Cet article propose un cadre hybride intégrant l'apprentissage automatique et des données économétriques pour prévoir la valeur des joueurs sur le marché, en s'appuyant sur les données d'événements à haute résolution de StatsBomb et sur l'historique des transferts. En synthétisant des mesures avancées (e.g., les xG) et des variables de marché externes, les résultats préliminaires indiquent une précision prédictive prometteuse.

Mots-clés. Valorisation des joueurs de football, Prévision de la valeur marchande, Apprentissage automatique, Analyse du marché des transferts, Modèles hybrides.

Abstract. The global football transfer market, a multi-billion-dollar ecosystem, demands accurate player valuation to navigate its economic complexities. Traditional methods, relying on heuristic negotiations and basic performance metrics, often fail to capture nonlinear interactions among factors such as on-field performance, demographic trends, contractual constraints, and macroeconomic dynamics. This paper proposes a hybrid framework integrating machine learning with econometric insights to forecast player market values, leveraging high-resolution event data from StatsBomb and historical transfer records. By synthesizing advanced metrics (e.g., expected threat, possession-adjusted actions) and external market variables, preliminary results indicate promising predictive accuracy.

Keywords. Football player valuation, Market value forecasting, Machine learning, Transfer market analytics, Hybrid models.

1 Introduction

The global football transfer market, valued at over $10 billion annually [1], represents a high-stakes economic ecosystem where accurate player valuation is critical for clubs, agents, and investors. Traditional valuation methods, rooted in heuristic negotiations and simplistic performance metrics (e.g., goals, assists), increasingly fail to capture the complexity of modern football dynamics. Market values are influenced by a confluence of factors—on-field performance [2], demographic trends [3], contractual constraints, and macroeconomic forces—yet existing models often overlook nonlinear interactions between these variables.

Recent advances in data availability (e.g., event-level tracking, transfer fee databases) and machine learning [4] present an opportunity to redefine valuation paradigms. This paper addresses the challenge of forecasting player market values by synthesizing high-resolution performance data with econometric insights, offering a framework that balances predictive accuracy with interpretability. Our work contributes to both sports analytics and decision science by demonstrating how hybrid models can mitigate biases inherent in human-driven valuations while adapting to the transfer market's inherent volatility.

2 Related Work

Forecasting football players' market value has evolved from heuristic club valuations to data-driven predictive modeling, drawing insights from econometrics, machine learning, and sports analytics. This section synthesizes foundational and contemporary approaches, highlighting methodological advancements and unresolved challenges.

Econometric Foundations. Early work by [3] established human capital theory in football, linking transfer fees to wages and club revenue. Their model, however, relied on static financial indicators, neglecting dynamic market forces. [1] identified performance metrics (goals, assists) and demographic factors (age, contract length) as value determinants using linear regression, but their linear assumptions overlooked nonlinear interactions.

Machine Learning and Feature Engineering. Recent advances leverage high-dimensional data. [2] pioneered the VAEP (Valuing Actions by Estimating Probabilities) framework, quantifying in-game actions' impact on match outcomes, which became a benchmark for player valuation features.

Gaps identified. Despite progress, three gaps persist: (1) underspecification of temporal dynamics (e.g., aging curves, contract expiration), (2) overreliance on structured data, ignoring unstructured sources (e.g., agent influence, media narratives), and (3) lack of interpretability in so-called "black-box" models.

3 Case Study - Forecasting Players Market Value

In this section we present a case study for forecasting player market value with StatsBomb advanced statistics and transfer data.

3.1 Problem Formulation

We frame market value forecasting as a supervised learning task, where the target variable y is a player's transfer fee (or estimated market value for non-transferred players), and the feature set $\mathbf{X}$ comprises covariates spanning four domains: *performance*, *demographic*, *contractual*, and *market-contextual* factors. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote the dataset, where each observation corresponds to a player-season pair. The objective is to learn a mapping:

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

3.2 Datasets

StatsBomb Event Data We utilize event-level data from StatsBomb [5], which provides granular spatiotemporal records of on-ball actions (e.g., passes, shots, dribbles) across 10,000+ matches. Many features are available including:

- Possession-adjusted metrics (e.g., defensive actions/90)
- Contextual performance (e.g., passes under pressure)
- Advanced statistics (e.g., xG)

Transfer Market Data Transfer fees and market valuations are sourced from the open-access repository [6], which aggregates historical transfers from the 90s to 2023 (see Figure 1). The dataset includes:

- Inflation-adjusted transfer fees (€ million)
- Contract start/end dates and wage data
- Club financial indicators (e.g., revenue, squad size)

4 Experimental Results

The final analysis of the ongoing experiments is still in progress. While the detailed findings are not yet available, we are pleased to confirm that the comprehensive results, including

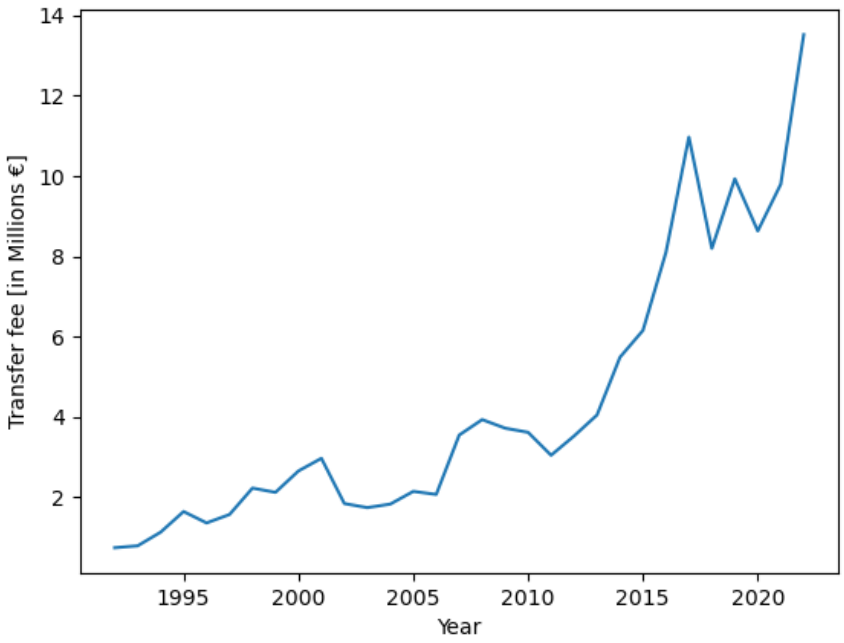


Figure 1: Evolution of transfer fees in the English Premier League

comparative evaluations of model performance, will be presented in full at the upcoming conference. We appreciate your patience and look forward to sharing insights that could contribute to a better understanding of football players' market value.

References

[1] R. Poli, L. Ravenel, and R. Besson, "Scientific evaluation of the transfer value of football players." Football Observatory Report, 2020.

[2] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, "Actions speak louder than goals: Valuing player actions in soccer," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1851–1861, 2019.

[3] E. Van den Berg, "The valuation of human capital in the football player transfer market," *Rotterdam: ErasmusUniversity*, 2011.

[4] K. Zatloukal, "Transfers and success of football players," 2023.

[5] StatsBomb, "Statsbomb open data." https://statsbomb.com, 2023.

[6] M. Ewen, "Football transfer data." GitHub Repository, 2023.

Proposition pour détecter des momentums dans un match

Christian Derquenne

Chercheur indépendant - chris.emr@wanadoo.fr

Résumé. Lors de matches d'opposition individuelle ou collective, des séquences gagnantes peuvent apparaitre de façon significative à une ou plusieurs reprises et plus ou moins réparties entre les adversaires. Ces séquences gagnantes sont généralement nommées "momentums". Par exemple, lors d'un match de tennis, un enchainement de points gagnants peut se produire et influencer ou non le résultat final. Nous proposons une démarche de recherche de momentum reposant sur les principes de détection de défauts dans un processus industriel en utilisant notamment la loi Géométrique. Nous appliquons cette démarche au match de la finale du tournoi de Roland Garros de 2020 opposant Raphael Nadal à Novak Djokovic.

Mots-clés. Matches d'opposition, Momentum, loi géométrique, test de detection

Abstract. During individual or collective opposition matches, winning sequences can appear significantly on one or more occasions and more or less distributed between the opponents. These winning sequences are generally called "momentums". For example, during a tennis match, a series of winning points can occur and influence or not the final result. We propose a momentum detection approach based on the principles of defect detection in an industrial process, using in particular the Geometric distribution. We apply this approach to the match of the final of the Roland Garros tournament of 2020 between Raphael Nadal and Novak Djokovic.

Keywords. Opposite games, Momentum, Geometric distribution, detection test

1 Contexte - objectif

Préambule : Cet article fait suite à une discussion et à un exposé de Léo et Louis Gerville-Réache lors de Rencontres de Statistique UBS de Vannes, en novembre 2024.

Dans de nombreux sports individuels ou collectifs comme le tennis ou le volley par exemple, nous pouvons observer au cours du match une série relativement longue de points gagnés ou perdus par un joueur ou une équipe. Cette série est mathématiquement une simple suite (ou séquence) binaire de 1 et de 0 où le chiffre 1 traduit la victoire et le 0 la perte du point.

Nous reprenons l'exemple sur lequel nous avons discuté : le match opposant Rafael Nadal à Novak Djokovic lors de la finale du tournoi de Roland Garros en 2020. Cette finale est remportée par Rafael Nadal sur le score de : 6/0 ; 6/2 ; 7/5, pour 183 points joués. La série de points gagnés (1) et perdus (0) par Raphael Nadal est la suivante :

0100111101110011011101101001010111000111101011101110111001001
0010011111001010110111001101101011111110000111001100000111011
0101000011011111110010100010001011010101011000011111001111111

Le problème formulé dans cet exposé était de détecter statistiquement des "moments forts" (séquences de points atypiques) et des "momentums" (périodes où un joueur surpasse les attentes théoriques). Pour répondre à ce problème il est nécessaire de construire un modèle mathématique de la probabilité de gain et de perte de chaque point de la rencontre, puis un momentum pourrait être vu comme une série "atypique" de points au sein de la longue série constituant le match. En fait, une série "anormalement" dense en 1 pourrait être définie comme un momentum en faveur de Rafael Nadal alors qu'une série "anormalement" dense en 0 serait un momentum en faveur de Novak Djokovic. Détecter un momentum consisterait donc à comparer la série binaire de la rencontre avec des probabilités "théoriques" associées à chaque point... La méthode proposée dans [Gerville-Reache et al., 2024] consiste tout d'abord à modéliser la probabilité de gagner chaque point en utilisant une adaptation du modèle de Volf et Kourim (2024) : un modèle logit booléen temporel avec des covariables. Les auteurs proposent une définition statistique originale du concept de momentum reposant d'une part, sur les probabilités de toutes les séries de points successifs, débutant au point $t, (t = 1, ..., T)$ et de longueur $k, (k = 1, ..., T - t + k)$, et d'autre part, sur les mêmes séries observées au cours du match. Ils contruisent un indicateur de longueur de séquences de points, alors la valeur de celui-ci à un moment t sera un moment fort si et seulement si la probabilité que le nombre de points marqués dans la séquence soit supérieur ou égal au nombre de points réellement marqués, est particulièrement faible (par exemple $\alpha = 0,05)$). Par la suite ils introduisent une approche d'apprentissage progressif pour prédire le résultat des points de la finale en temps réel. Cette procédure consiste à entraîner de manière itérative un modèle en intégrant progressivement de nouvelles observations au fur et à mesure qu'elles se produisent. Cette technique est particulièrement pertinente dans un contexte sportif, où le résultat de chaque point peut fournir des informations supplémentaires pour affiner les prédictions futures.

Nous proposons une autre approche pour détecter des momentums fondée sur la loi Géométrique utilisée dans des contextes industriels. Puis nous appliquons cette approche sur l'exemple introductif et nous comparons les résultats à ceux de l'article décrit précédement. Nous concluons sur les apports et faiblesses de cette nouvelle proposition, l'utilisation dans d'autres domaines d'applications et nous fournissons quelques voies futures en termes d'amélioration.

2 Une nouvelle proposition et son extension

2.1 Une première idée

Ce type de séquence fait penser à la détection de défauts dans un processus industriel. Ce dernier est généralement modélisé par une loi Géométrique formalisée de la façon suivante : soit une séquence d'essais issus de lois de Bernoulli indépendantes, où la probabilité de succès à chaque essai est égale à $p(0 < p < 1)$ $(q = 1 - p)$, la réalisation issue d'une loi Géométrique $\mathcal{G}(p)$ correspond au nombre d'essais ou de défauts n avant le premier succès.

Densité de probabilité $f_{\mathcal{G}}(q) = pq^n$ et ***Fonction de répartition*** $F_{\mathcal{G}}(q) = 1 - q^{n+1}$

Généralement, on suppose que la probabilité p d'apparition d'un défaut est fixée par l'expert du processus industriel. Ici, la probabilité de succès ou d'échec évolue au cours du match et peut être

vue comme une succession de lois de Bernoulli $X(t)$ non identiquement distribuées. On supposera cependant qu'elles sont indépendantes. Les probabilités successives au cours du temps sont : $(p(1), ..., p(t), ..., p(T))$ où T correspond à la longueur totale de la séquence et, $p(t) = (n_1(t))/t$ désigne la probabilité de succès au temps t et $n_1(t)$ est le nombre de succès jusqu'au temps t du joueur 1. A l'opposé, pour le joueur 2, nous avons : $q(t) = (n_2(t))/t = 1 - p(t)$.

Par conséquent, un momentum sera susceptible d'intervenir s'il y a un nombre suffisant de succès consécutifs au regard de l'évolution des probabilités de succès. Pour tester l'apparition d'un momentum, nous fixons les hypothèses nulle et alternative telles que :

$\mathbf{H}_0$: absence d'un momentum au temps t *vs* $\mathbf{H}_1$: présence d'un momentum au temps t

La statistique de test est le nombre de points successifs : $N_1(t)$ suit, sous $\mathbf{H}_0$, une loi Géométrique de paramètre $p_1 = p(t_0)$ où $p(t_0)$ désigne la probabilité de succès du joueur 1 juste avant le début de sa nouvelle séquence de succès et le temps $t_0 = (t-1) \times 1_{[x(t-1)=0]}$ avec $x(t-1)$ la réalisation d'une loi de Bernoulli égale à 0 pour le joueur 1 car le joueur 2 a remporté le dernier point, pour lui : $y(t-1) = 1$. Le choix de p_1 est justifié car il tient compte du passé immédiat (juste avant le premier succès du joueur 1), mais pas d'un passé dynamique. Cette solution pourrait être mise en oeuvre. La p-valeur : $pval_1$=Pr$[\mathcal{G}(p(t)) > n_1(t)] = p_1^{(n_1(t))}$ où $n_1(t)$ est le nombre observé de succès à partir du temps $t_0 + 1$. Les p-valeurs successives vont mécaniquement diminuer jusqu'à un nouvel échec du joueur 1. On nomme $n_1^* = \max_t n_1(t)$ correspond au dernier nombre de succès obtenu par le joueur 1 avant son échec suivant, face au joueur 2.

La règle de décision pour déterminer la présence d'un momentum est la suivante : si $p_1^{(n_1^*)} < \alpha$ alors la séquence de succès observée au temps t est un momentum. Dès que cette inégalité est observée, le joueur est dans un momentum. En d'autres termes, plusieurs p-valeurs peuvent être inférieures à α dès que le momentum a débuté et se poursuivre encore un certain temps. Il est également possible de calculer le fractile de la loi Géométrique pour décider le rejet de l'hypothèse nulle. Pour cela, $F_{\mathcal{G}(p_1)}(n_1(t))$=Pr$[\mathcal{G}(p_1(t)) < n_1(t)] = 1 - p_1^{n1(t)}$ représente la fonction de répartition de $\mathcal{G}$, alors son fractile $1 - \alpha$ est égal à $F^{-1}_{\mathcal{G}(p_1)}(1-\alpha) = (ln(\alpha))/(ln(p_1) = n_\alpha(p_1))$. Par exemple, si $p_1 = 0,6$ et $\alpha = 0,05$, alors $n_\alpha = 5,86$ arrondie à 6. Par conséquent au seuil α, une séquence significative de succès devra avoir une longueur supérieure ou égale à 6.

Remarque importante : le nombre nécessaire de gains successifs pour obtenir un momentum peut être différent entre les deux joueurs. En effet, plus un joueur possède une probabilité forte de gagner face à son adversaire au fil du match, moins l'apparition d'un momentum sera probable, donc observé.

2.2 Une seconde idée : extension à la proposition

Un momentum peut également être composé de plusieurs séquences significatives de gains séparées par de toutes petites séquences d'échecs. Comment mettre en oeuvre cette remarque ?

Soit $s_i(t_i, t_{i+1} - 1)$, la i^{ieme} séquence au cours du match, de longueur $t_{i+1} - t_i$ notée $L_i^{(obs)}$. Tout d'abord, celle-ci est équivalente à n_1^*, le dernier nombre de succès obtenu par le joueur 1 avant son échec suivant face au joueur 2. Nous savons également que si $p_1^{n_1^*} < \alpha$ alors la séquence de succès observée au temps t est un momentum, où p_1 désigne la probabilité de succès du joueur 1 juste avant le début de sa nouvelle séquence de succès.

L'objectif est de détecter l'existence d'un momentum dans le cas d'une suite de séquences dont certaines peuvent être non significatives, comme définies plus haut à l'aide de la loi Géométrique.

Plaçons-nous dans le cas où un joueur enchaine les trois séquences : s_{i-1}, s_i, s_{i+1}, de longueurs respectives : $L_{i-1}^{(obs)}, L_i^{(obs)}, L_{i+1}^{(obs)}$. La première et la troisième séquences correspondent à des successions de succès du joueur 1, alors que la deuxième représente des successions d'échecs.

Pour réaliser le test de détection, il est nécessaire que l'une des deux séquences de succès du joueur 1 soit significative au sens du test avec la loi Géométrique. Si $L_i^{(obs)}$ est suffisamment petite au regard des deux autres longueurs de séquences alors celles-ci formeront un momentum.

Par ailleurs, nous définissons trois nouveaux types de momentum :

Momentum-avant composé de trois séquences : la première contient des succès non significatifs, la deuxième possède des échecs dont la longueur fera l'objet du test de détection, une troisième validée comme momemtum à l'aide du test avec la loi Géométrique ;

Momentum-après composé de trois séquences également : la première correspond à des succès validés comme momemtum, une deuxième se caractérise par des échecs qui fera aussi l'objet du test de détection et la troisième est relative à des succès non significatifs ;

Momentum-avant-après composé de cinq séquences mêlant celles des deux types de momentums précédents : la première contient des succès non significatifs, la deuxième possède des échecs à valider ou non avec le test, la troisième est validée comme momemtum, la quatrième se caractérise par des échecs à tester et la cinquième est relative à des succès non significatifs.

La procédure de test se déroule comme suit.

L'hypothèse nulle correspond à l'absence d'un prolongement d'un momentum en un des trois types de momentums introduits précédemment contre l'hypothèse alternative offrant un momentum prolongé. Sous cette hypothèse, L suit une loi Binomiale : $\mathcal{B}(n, p)$ où n est la somme des trois longueurs : $L_{i-1}^{(obs)}, L_i^{(obs)}, L_{i+1}^{(obs)}$ et $p = 1 - p_1 = q_1$ qui désigne la probabilité d'échec du joueur 1 juste avant le début de sa séquence de succès s_{i-1}.

Construction de la statistique de test :

Pour un momentum-avant, la séquence s_{i-1} de longueur $L_{i-1}^{(obs)}$ représente des succès non significatifs, la séquence s_i de longueur $L_i^{(obs)}$ correspond à des échecs (c'est la statistique de test) et la séquence s_{i+1} de longueur $L_{i+1}^{(obs)}$ est relative à un momentum ;

Pour un momentum-après, la séquence s_{i-1} de longueur $L_{i-1}^{(obs)}$ représente un momentum, la séquence s_i de longueur $L_i^{(obs)}$ correspond à des échecs (c'est la statistique de test) et la séquence s_{i+1} de longueur $L_{i+1}^{(obs)}$ est relative à des succès non significatifs.

Pour un momentum-avant-après, si les momentums avant et après sont significatifs alors un nouveau test est effectué en tenant compte des cinq séquences : première s_{i-2} de longueur $L_{i-2)}^{(obs)}$ représente des succès non significatifs, la séquence s_{i-1} de longueur $L_{i-1}^{(obs)}$ correspond à des échecs, la séquence s_i de longueur $L_i^{(obs)}$ est relative à un momentum, la séquence s_{i+1} de longueur $L_{i+1}^{(obs)}$ correspond à des échecs et la séquence s_{i+2} de longueur $L_{i+2}^{(obs)}$ est relative à des succès non significatifs. La statistique de test est la somme : $L_{i-1}^{(obs)} + L_{i+1}^{(obs)}$.

Pour les deux premiers tests, la règle de décision est la suivante :

si $\Pr[L \leq L_i^{(obs)}] = \sum_{l=0}^{L_i^{(obs)}} \binom{l}{n}(1-p_1)^l p_1^{n-l} < \alpha$ alors, on rejette l'hypothèse nulle, où p_1 est la probabilité d'échec du joueur 1 ; il y a présence d'un momentum prolongé avant ou après.

Pour le troisième test, nous calculons une nouvelle probabilité pour deux échecs du joueur 1, $p_{1+} = \frac{(L_{i-2}^{(obs)}+L_{i-1}^{(obs)}+L_i^{(obs)})p_{1(i-3)}+(L_i^{(obs)}+L_{i+1}^{(obs)}+L_{i+2}^{(obs)})p_{1(i-1)})}{L_{i-2}^{(obs)}+L_{i-1}^{(obs)}+L_i^{(obs)}+L_{i+1}^{(obs)}+L_{i+2}^{(obs)}}$ pour la loi Binomiale sous l'hypothèse nulle : $\mathcal{B}(n,p)$, où $p = 1-p_{1+}$ et $n = L_{i-2}^{(obs)} + L_{i-1}^{(obs)} + L_i^{(obs)} + L_{i+1}^{(obs)} + L_{i+2}^{(obs)}$.

La règle de décision est la suivante :

si $\Pr[L \leq L_{i-1}^{(obs)} + L_{i+1}^{(obs)}] = \sum_{l=0}^{L_{i-1}^{(obs)}+L_{i+1}^{(obs)}} \binom{l}{n}(1-p_{1+})^l p_{1+}^{n-l} < \alpha$ alors on rejette l'hypothèse nulle: il y a présence d'un momentum prolongé avant et après.

2.3 Application au match Nadal *vs* Djokovic

La figure 1 visualise l'évolution temporelle des points des deux joueurs. Les séquences vertes correspondent à des succès de Nadal ; les rouges représentent les succès de Djokovic. La figure 2 affiche en rouge l'évolution du nombre de points marqués par Nadal et en bleu l'évolution des fractiles seulement à la fin de chaque momentum (il y en a quatre). Nous pouvons constater que chaque nombre de succès est supérieur à son fractile associé (quatre sommets rouges sont au-dessus des sommets bleus). Il en va de même pour Djokovic qui a quatre momentums détectés (fig. 3), les pointes vertes dépassent les pointes bleues. La figure 4 affiche la présence de quatre momentums pour chacun des deux adversaires (en vert pour Nadal ; en bleu pour Djokovic).

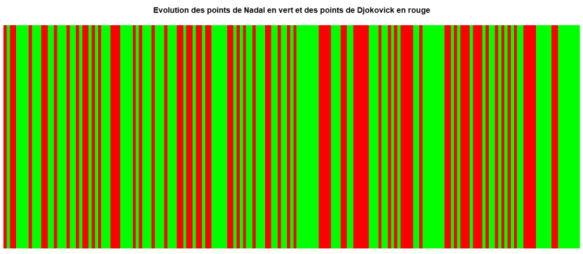


Figure 1 : Evolution des points de Nadal vs Djokovic

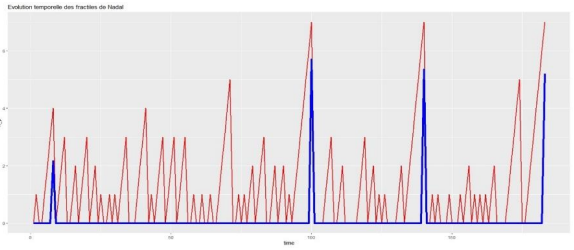


Figure 2 : Evolution des fractiles/nombres des succès Nadal

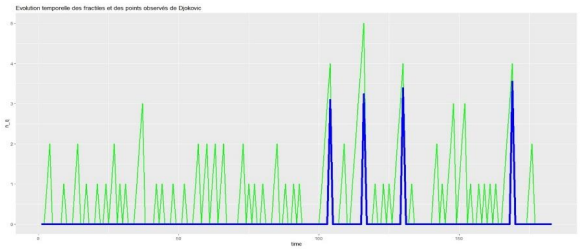


Figure 3 : Evolution des fractiles/nombres des succès Djoko

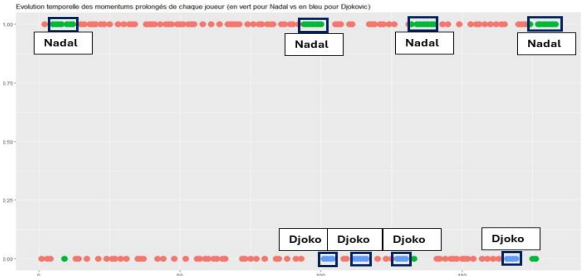


Figure 4 : Evolution des momentums prolongés

Le premier de Nadal est de type momentum-après (cf. tab. 1). La p-valeur du momentum courant ($t = 5, 8$) de Nadal vaut 0,004, et celle de sa prolongation après ($t = 5, 12$) est égale

à 0,001. Remarquons que le test du momentum-avant n'est pas significatif (p-valeur=0,376). Signalons que comme ce momentum se situe au début du match ; le calcul de la probabilité de succès de Nadal risque d'être incertaine. Le deuxième momentum de Nadal n'a pas été prolongé, les troisième et quatrième momentums sont de type momentum-avant. Djokovic ne possède pas de momentum prolongé. Dans la méthode [Gerville et al., 2024], le premier momentum n'est pas détecté, mais cela peut être logique car il est au début du match. Le deuxième est prolongé avant ($t = 86, 93 + t = 94, 100$) alors que nous ne nous l'avons pas étendu ; le troisième est prolongé après ($t = 134, 140 + t = 141, 154$), alors que nous l'avons étendu avant et le quatrième n'est pas prolongé, alors que nous l'avons étendu avant.

Joueur	**p-val(av)**	**p-val(ap)**	**p-val(co)**	**Avant**	**Après**	**Courant**
Nadal	0,376	**0,001**	**0,004**	0	**5-12**	**5-8**
Nadal	0,065	0,258	**0,025**	0	0	**94-100**
Djokovic	1,000	1,000	**0,021**	0	0	**101-104**
Djokovic	1,000	1,000	**0,010**	0	0	**112-116**
Djokovic	1,000	1,000	**0,029**	0	0	**127-130**
Djokovic	**0,031**	0,126	**0,020**	**132-140**	0	**134-140**
Nadal	1,000	1,000	**0,034**	0	0	**166-169**
Nadal	**0,019**	1,000	**0,018**	**174-183**	0	**177-184**

Table 1: Evolution des 8 momentums prolongés de chaque joueur

3 Synthèse des résultats et voies futures

Cet article propose une nouvelle approche pour détecter des momentums dans des matches opposant deux adversaires individuels. Elle repose sur une loi Géométrique généralement utilisée dans des processus industriels pour la détection de défauts. Cette méthode a été étendue à l'aide d'une loi Binomiale pour détecter l'existence d'un éventuel prolongement avant et/ou après le momentum courant. Les résultats obtenus sur la finale Nadal/Djokovic sont tout à fait encourageants et sont comparables à ceux de Gerville et al. Si l'application de ces approches parait directe au volley-ball car il y a le même type de séquences 0/1 de points perdus et gagnés. Leur adaptation à d'autres sports d'opposition peut se révéler moins naturelle. Par exemple, au basket, la possession d'une équipe peut se conclure de différentes façons (panier réussi à 2 ou 3 points ou raté, faute commise/subie, lancer franc, sortie du ballon...). Dans les voies futures, nous travaillerons sur le calcul de la probabilité de succès avant le départ d'une série de points gagnés pendant le match et sur la première probabilité de succès au début du match, nous tiendrons compte du service et nous tenterons d'adapter notre méthode à d'autres sports d'opposition.

Bibliographie

Gerville-Réache L., Gerville-Réache L., Terret C. & Morlier J., (2024), Momentum en sport: approche statistique et application au tennis, *9ièmes Rencontres de Statistique de Vannes - Science des données du sport*, UBS - Vannes, France.
Volf P. & Kourim T. (2024), A model and application of binary random sequence with probabilities depending on history, *KYBERNETIKA*, Vol **60**, p.110-124.

SUR LA DÉTECTION STATISTIQUE DU MOMENTUM EN SPORT

Louis Gerville-Réache[1], Léo Gerville-Réache[2], Cédric Terret[2] et Julien Morlier[2]

[1] *ENSAI, Rennes, France, louisgervillereache@hotmail.com*
[2] *Univ. Bordeaux, IMS, UMR 5218, F-33400, Talence, France*
leo.gerville-reache@u-bordeaux.fr , cedric.terret@u-bordeaux.fr , julien.morlier@u-bordeaux.fr

Résumé. Cette communication se concentre sur la détection statistique du momentum en sport, en prenant l'exemple de la finale de Roland-Garros 2020 entre Rafael Nadal et Novak Djokovic. Le momentum désigne une dynamique positive ou négative dans la performance d'un joueur, observable par une série de points gagnés ou perdus. L'étude modélise cette dynamique en représentant les résultats de chaque point par une séquence binaire, où "1" indique une victoire et "0" une défaite. L'objectif est de détecter statistiquement des "moments forts" (séquences de points atypiques) et des "momentums" (périodes où un joueur surpasse les attentes théoriques). Un modèle mathématique, basé sur les probabilités conditionnelles des points successifs, est utilisé pour prédire la probabilité de victoire de chaque point en fonction de l'historique des matchs des deux joueurs. Une définition statistique originale des concepts de moment fort puis, finalement, de momentum est proposée. Les résultats montrent comment cette approche statistique permet d'identifier les phases de momentum, offrant ainsi une analyse statistique de la dynamique de la rencontre. Cette méthode pourrait être généralisée à d'autres sports, offrant ainsi de nouvelles perspectives dans l'analyse temporelle des rencontres sportives.

Mots-clés. Momentum, Séquence binaire, Dynamique de la performance sportive, Probabilité de victoire.

Abstract. This presentation focuses on the statistical detection of momentum in sports, using the example of the 2020 Roland-Garros final between Rafael Nadal and Novak Djokovic. Momentum refers to a positive or negative dynamic in a player's performance, observable through a series of points won or lost. The study models this dynamic by representing the results of each point as a binary sequence, where "1" indicates a win and "0" a loss. The goal is to statistically detect "strong moments" (sequences of atypical points) and "momentums" (periods where a player exceeds theoretical expectations). A mathematical model, based on the conditional probabilities of successive points, is used to predict the probability of winning each point based on the historical data of the two players' matches. An original statistical definition of the concepts of strong moment and, ultimately, momentum is proposed. The results demonstrate how this statistical approach helps identify momentum phases, thus offering a statistical analysis of the dynamics of the match. This method could be generalized to other sports, providing new perspectives in the temporal analysis of sporting events.

Keywords. Momentum, Binary sequence, Sport performance dynamics, Probability of winning

1 Introduction

Le concept de momentum en sport désigne, le plus souvent, l'élan ou la dynamique que peut prendre une équipe ou un athlète au cours d'une compétition. Ce terme est de plus en plus utilisé pour décrire une période durant laquelle une équipe ou un joueur semble être dans une phase de réussite, où tout fonctionne en leur faveur. Le momentum peut se manifester par une série de points marqués, une défense particulièrement efficace, ou encore par un changement d'attitude et de confiance. Pour Michael Kent (2006) le momentum sportif est : *Un changement positif ou négatif dans la cognition, l'affect, la physiologie et le comportement causé par un événement ou une série d'événements qui affecte soit les perceptions des compétiteurs, soit, peut-être, la qualité de la performance et le résultat de la compétition. Le momentum positif est associé à des périodes de compétition, telles qu'une série de victoires, où tout semble "bien se passer" pour les compétiteurs. En revanche, le momentum négatif est associé à des périodes, telles qu'une série de défaites, où tout semble "mal tourner".*

Dans nombres de sports comme le tennis ou le volley par exemple, on observe au cours d'un match une série relativement longue de points gagnés ou perdus. Cette série est mathématiquement une simple suite (ou séquence) binaire de 1 et de 0 où le chiffre 1 traduit la victoire et le 0 la perte du point. Voici la série de victoire et de perte de points de Rafael Nadal lors de sa finale à Roland Garros en 2020 contre Novak Djokovic. Cette finale est remportée par Rafael Nadal sur le score de 6/0-6/2-7/5 et 183 points ont été joués.

010011110111001101110110100101011100011110101110111011100100100100111100101011011100110110101111111000011100110000011101101010000110111111100101000100010110101010110000011111001111111

Comment détecter sur cette séquence un ou plusieurs momentums au sein de cette série ? Cette question nécessite, dans un premier temps, de construire un modèle mathématique de la probabilité de gain et de perte de chaque point de la rencontre. Ensuite, un momentum pourrait être vu comme une série « atypique » de points au sein de la longue série constituant le match. Plus précisément, une série « anormalement » dense en 1 pourrait être vu comme un momentum en faveur de Rafael Nadal alors qu'une série « anormalement » dense en 0 serait un momentum en faveur de Novak Djokovic. Détecter un momentum consisterait donc à comparer la série binaire de la rencontre avec des probabilités « théoriques » associés à chaque point.

Dans ce papier, nous proposons dans un premier temps de modéliser la probabilité de victoire de chaque point en utilisant une adaptation du modèle de Volf et Kourim (2024). Ensuite, nous proposons une définition statistique originale du concept de momentum basée d'une part sur les probabilités de l'ensemble des séries de points successifs, commençant au point i, $(i = 1, \dots, T)$ et de longueur k, $(k = 1, \dots, T - i + 1)$, et d'autre part les mêmes séries observées lors de la rencontre.

2 Modélisation

La modélisation des séquences binaires avec des probabilités dépendantes de l'historique a été étudiée dans divers contextes sportifs. Ces modèles permettent de capturer la dynamique des événements successifs, en tenant compte de l'influence des événements passés sur la probabilité des événements futurs. Les travaux de Albert et Koning (2007) ont utilisé des modèles de séquences binaires pour analyser les résultats des tirs au but dans le football, montrant que les probabilités des tirs réussis dépendent fortement du résultat des tirs précédents.

Le tennis, avec sa structure séquentielle distincte (alternance des services en fonction du

comptage spécifique des points), se prête particulièrement bien à l'application de modèles de séquence binaire. Klaassen et Magnus (2001) ont été parmi les premiers à modéliser les probabilités des points dans un match de tennis en prenant en compte les états précédents pour prédire les résultats futurs. Leur modèle markovien a démontré que la probabilité de gagner un point dépendait non seulement de la situation actuelle, mais aussi de l'historique des points précédents.

En 2024, Dans leur article *A Model and Application of Binary Random Sequence with Probabilities Depending on History*, Petr Volf et Tomas Kourim proposent une approche novatrice pour modéliser les séquences binaires avec des probabilités dépendantes de l'historique. Leur modèle utilise une régression logistique pour intégrer les covariables et les états historiques, offrant une méthode flexible pour analyser les séquences d'événements binaires. Ce modèle est particulièrement pertinent, car il permet de prendre en compte des variables telles que le résultat du point précédent, le serveur ou encore l'adversaire.

La modélisation de la série binaire des points la finale de Roland Garros 2020, se déroulent en trois temps : d'une part, nous modélisons la probabilité que Rafael Nadal gagne un point lors de la finale en utilisant l'ensemble de l'historique de ses rencontres lors du tournoi. Parallèlement, nous réalisons une modélisation similaire sur l'ensemble des rencontres de Novak Djokovic sur ce même tournoi. Enfin, nous construisons un modèle « hybride » spécifique à la finale en utilisant les deux modèles individuels précédents.

2.1 Les datas

De manière générique, pour un joueur de tennis de référence, les données utilisées sont de la forme suivante (Figure 1.):

Vainqueur du point	***1 ≡ Joueur de référence | 0 ≡ Adversaire***
Numéro du point dans le match	*1, 2, 3 …*
Nombre de points déjà gagnés (par le joueur de référence)	*1, 2, 3 …*
Serveur	*1 ≡ Joueur de référence | 0 ≡ Adversaire*
Adversaire	*Adversaire 1, Adversaire 2, …*

Figure 1. Modèle de données

Les données des matchs de Roland Garros 2020 joués par Rafael Nadal d'une part et Novak Djokovic d'autre part ont été recueillies sous la forme de feuilles de match à partir du site Github https://github.com/JeffSackmann/tennis_slam_pointbypoint/. Ces données comprennent chaque point joué pendant le match, enregistrant qui a servi, qui était l'adversaire et qui a gagné le point. Le modèle a donc été ajusté aux données spécifiques pour évaluer l'impact des variables explicatives (Numéro du point dans le match, Nombre de points déjà gagnés, serveur et Adversaire). Visuellement (Figure 2.), la finale peut se représenter sous la forme de l'alternance des points gagnés (en vert par Novak Djokovic et orange pour Rafael Nadal) sur la première ligne. La seconde ligne représentant l'alternance de services.

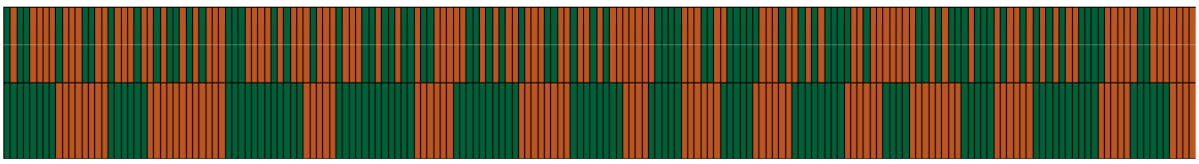


Figure 2. Séquence de Points et de Services de la Finale de Roland Garros 2020

Cette représentation est générique dans le sens où elle peut être utilisée pour de nombreux autres sport comme le volley, le tennis de table, le badminton, le handball… Dans le but d'une généralisation à de nombreuses autres situations d'opposition sportive, les spécificités du score en jeux et sets au tennis n'ont pas été pris en compte explicitement.

2.2 Le modèle logit temporel avec covariables

Le modèle décrit par Petr Volf et Tomas Kourim en 2024 est une extension de la régression logistique pour modéliser des séquences binaires avec des covariables dépendant du temps. Le modèle considère une séquence de variables aléatoires binaires X_t, prenant les valeurs 1 ou 0 (on parle également de processus stochastique binaire à temps discret). Ces variables représentent les résultats possibles de points successifs dans un processus temporel. Il est supposé que les probabilités $P_t = \mathbb{P}(X_t = 1 \mid \mu_t, \boldsymbol{z}(t))$ suivent un modèle de régression logistique, où le logit de la probabilité suit la forme suivante :

$$logit(P_t) = \mu_t + \boldsymbol{\beta}'\boldsymbol{z}(t) \quad \text{avec} \quad logit(P_t) = ln\left(\frac{P_t}{1-P_t}\right)$$

μ_t est la composante temporelle.
$\boldsymbol{\beta}$ est le vecteur des paramètres de la régression logistique.
$\boldsymbol{z}(t)$ est le vecteur des covariables observées au temps t.

Au début de la séquence ($t = 1$), μ_1 est initialisée à une valeur de départ. En l'occurrence ce sera ici la constante regroupant les modalités de référence des covariables catégorielles. Pour les instants suivants (t = 2, 3, …), μ_t est mis à jour en fonction des résultats des points précédents. La mise à jour de μ_t dépend du type de séquence binaire observée. Pour une séquence donnée, μ_t peut être mis à jour selon la formule générale :

$$\mu_{t+1} = \mu_t + c_1 x_t + c_2(1 - x_t) = \mu_t + c_2 + x_t(c_1 - c_2) = \mu_1 + tc_2 + (c_1 - c_2)\sum_{j=1}^{t} x_j$$

Où c_1 et $c_2$sont des coefficients qui influencent l'évolution de μ_t en réponse aux résultats des points précédents, et x_t est l'issue du point à l'instant t (la réalisation de la variable de Bernoulli X_t).

L'estimation des paramètres du modèle se fait par maximum de vraisemblance. La fonction de vraisemblance pour une seule séquence de temps $t = 1, 2, \ldots, T$ est définie en utilisant la forme logistique des probabilités conditionnelles P_t.

$$\mathcal{L} = \prod_{t=1}^{T} P_t^{x_t} \cdot (1 - P_t)^{(1-x_t)} = \prod_{t=1}^{T} \frac{\exp\left[\left(\mu_t + \boldsymbol{\beta}'\boldsymbol{z}(t)\right)x_t\right]}{1 + \exp\left(\mu_t + \boldsymbol{\beta}'\boldsymbol{z}(t)\right)}$$

Dans le cadre où N matchs sont observés et en notant $X_{t,n}, t = 1, \ldots, T_n, n = 1, \ldots, N$ le $t^{\text{ème}}$ point du $n^{\text{ème}}$ match, la vraisemblance s'écrit :

$$\mathcal{L}_N = \prod_{n=1}^{N} \prod_{t=1}^{T_n} \frac{\exp\left[\left(\mu_t + \boldsymbol{\beta}'\boldsymbol{z}_n(t)\right)x_t\right]}{1 + \exp\left(\mu_t + \boldsymbol{\beta}'\boldsymbol{z}_n(t)\right)}$$

L'algorithme de Newton-Raphson est utilisé pour calculer les estimations des paramètres et les matrices de variances-covariances asymptotiques (fonction glm du logiciel R).

2.3 Moment fort et momentum

Une séquence de point commençant au temps i et de longueur k sera un *moment fort* au niveau α selon la définition suivante :

$$MF(\alpha,i,k)=\begin{cases}[\![i\ , \quad i+k-1]\!] \text{ si } p(i,k)<\alpha \\ \emptyset \qquad\qquad\qquad\quad \text{sinon}\end{cases} \quad \text{avec} \quad p(i,k)=\mathbb{P}\left(\sum_{j=i}^{i+k-1} X_j \geq \sum_{j=i}^{i+k-1} x_j\right)$$

$p(i,k)$ est la probabilité, calculée par le modèle, que Rafael Nadal (par exemple ici) gagne au moins autant de points que ce qu'il a effectivement gagné au cours de la série de longueur k commençant au point i.

Aussi, $MF(\alpha,i,k)$ sera un moment fort si et seulement la probabilité, que le nombre de points marqués sur la séquence soit supérieure ou égale au nombre de points effectivement marqués, soit particulièrement petite.

On définit également moment fort de niveau α commençant au point i :

$$MF(\alpha,i)=\bigcup_{k\in[\![1\ ,\ T-i+1]\!]} MF(\alpha,i,k)$$

On définit également l'ensemble des points du match appartenant à l'union de tous les moments forts $MF(\alpha,i)$ pour tout $i\in[\![1\ ,\ T]\!]$:

$$MF(\alpha)=\bigcup_{i\in[\![1\ ,\ T]\!]} MF(\alpha,i)$$

Enfin, un momentum $M(\alpha,i)$ de niveau α commençant en i est défini comme le plus grand ensemble consécutif de points dont chaque point appartient à un moment fort.

$$M(\alpha,i)=\begin{cases}[\![i\ ,\ i+card\big(MF(\alpha,i)\big)-1]\!] \text{ si } i-1\notin MF(\alpha) \text{ et } i\in MF(\alpha) \\ \emptyset \qquad\qquad\qquad\qquad\qquad\quad \text{sinon}\end{cases}$$

3 Résultat

Les modélisations réalisées et l'application de la définition des moments forts et momentums sont décomposée selon le schéma suivant : Dans un premier temps, les parcours de Rafael Nadal d'une part puis de Novak Djokovic d'autre part sont modélisées afin de construire un modèle « hybride » des probabilités de victoire de chaque point concernant la finale. Fort de cette modélisation, l'algorithme de détection des moments fort puis des momentums est alors appliqué. La Figure 3 présente les résultats principaux. Nous y retrouvons les probabilités de victoire de chaque point pour Rafael Nadal, telles que le modèle présenté en partie 2 le propose, les fréquences de points gagnés par jeu et les différents moments forts ainsi que les momentums détectés.

Au niveau $\alpha=0{,}025$, nous détection ici trois momentums pour Rafael Nadal (6 ou 7 points consécutifs) et un pour Novak Djokovic. Pour ce dernier, s'agit d'une séquence plus complexe qui n'est constitué que d'un seul moment fort (p_Djokovic (141,14)=0.024). Sur ce momentum, Novak Djokovic gagne 10 points sur 14.

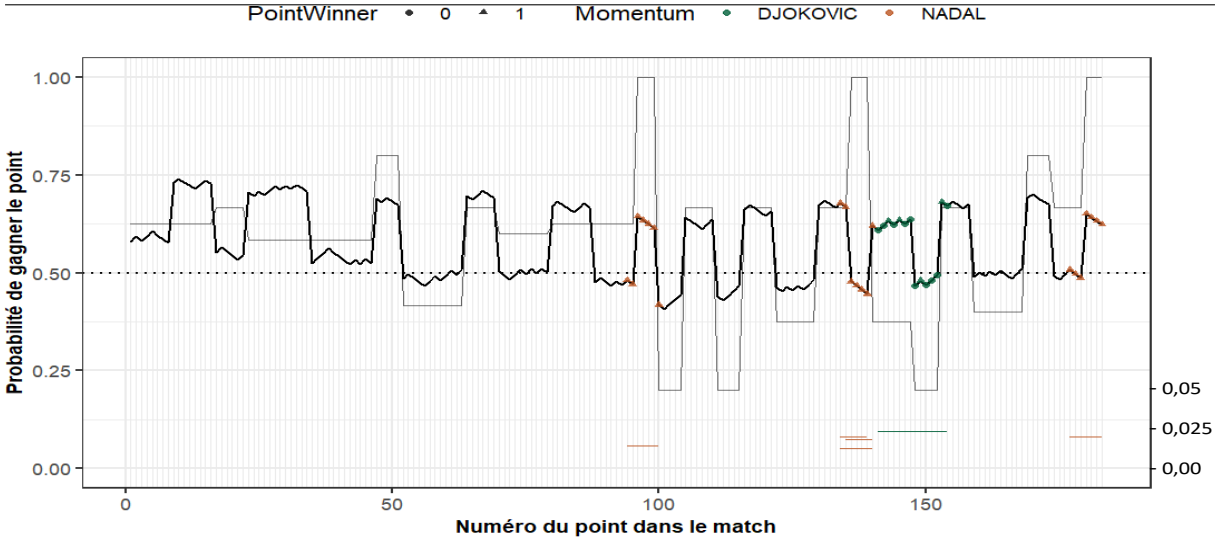


Figure 3. Momentums $M(\alpha, i)$ et moments forts $MF(\alpha, i, k)$ au niveau 0.025 de la finale Roland Garros 2020

4 Discussion

La détection de momentums est une question sportive complexe. Les causes d'une série de points atypiques sont généralement multiples et difficiles à objectiver. Dans cet article, nous avons cherché une approche statistique de détection de déséquilibres temporaires et inattendus du rapport de force en se basant essentiellement sur la modélisation de la série binaire des points de la rencontre (a posteriori mais aussi réalisable en temps réel). Selon notre approche, le concept de momentum est associé à un niveau α donné. Le choix de ce niveau est, comme souvent en statistique, une question délicate. Pour autant, une simulation numérique permet de calculer une p-value pour chaque momentum détecté de niveau α. Cette p-value simulée prend en compte la position du premier point (i) du momentum dans la série de points, la longueur du momentum (k) et la probabilité $p(i, k)$.

Bibliographie

Albert J. and Koning R., (2007). *Statistical Thinking in Sports* (1st ed.). CRC Press.

Briki W., Den Hatigh R.J.R. and Gernigon C. (2016), Le momentum psychologique en sport : vers une perspective complexe et dynamique, *Psychologie Française* Vol 61, p.91-102.

Klaassen F.J.G.M. and Magnus J.R. (2001), Are points in tennis independent and identically distributed ? Evidence from a dynamic binary panel data model, *Journal of the American Statistical Association* 96, p.500-509.

Kent M. (2006). *Oxford dictionary of sports science and medicine* (3rd ed.). New York: Oxford University Press.

Kourim T. (2019), Random walks with memory applied to grand slam tennis matches modeling, *MathSport International 2019: Conference Proceedings Propobos Publications*.

Ruscio J. and Brady K. (2019), Critical thinking as scientific reasoning ; examining the power of sports momentum, *The College of New Jersey*, Chapter 12.

Steeger, G.M. Dulin J.L. and Gonzalez G.O. (2021), Winning and losing streaks in the National Hockey League: are teams experiencing momentum or are games a sequence of random events ?, *Journal of Quantitative Analysis in Sports*, Volume 17 Issue 3.

Volf P. and Kourim T. (2024), A model and application of binary random sequence with probabilities depending on history, *KYBERNETIKA*, Volume 60, p.110-124.

Statistique mathématique

Pour détecter une rupture faible dans la moyenne conditionnelle d'un modèle hétéroscédastique non-linéaire

Fatma Aouissaoui [1,2] & Joseph Ngatchou-Wandji [1,3] & Hamdi Fthallah[2]

[1] *Université de Lorraine, Institut Élie Cartan de Lorraine, France,*
[2] *Université de Sousse,Laboratoire LAMMDA, École Supérieure des Sciences et de Technologie de Hammam Sousse, Tunisie,*
[3] *Université de Rennes (EHESP), Institut Élie Cartan de Lorraine, Nancy, France*

E-mails: aouissaouifatma4@gmail.com, joseph.ngatchou-wandji@univ-lorraine.fr, hamdi.fathallah@essths.u-sousse.tn

Résumé. Nous proposons un test de rapport de vraisemblance (RV) pour détecter la présence d'une rupture faible dans la moyenne conditionnelle d'une classe de modèles CHARN. Le comportement asymptotique de la statistique de test est étudié sous l'hypothèse nulle d'absence de rupture, et sous une suite d'alternatives locales. Des simulations numériques montrent que notre test a de bonnes performances sur les exemples considérés.

Mots-clés. Rupture, test du rapport de vraisemblance, propriété LAN, convergence faible, processus stochastique.

Abstract. We study a likelihood ratio (LR) test for detecting the existence of a weak change in the conditional mean of CHARN models. The asymptotic behavior of the test statistic is studied under the null hypothesis of no change, and under a sequence of local alternatives of a weak change. A simulation experiment shows that our test performs well on the examples tested.

Keywords. Change point, likelihood ratio test, LAN property, weak convergence, stochastic process.

1 Introduction

Le but de ce travail est l'étude de la détection des petites ruptures dans la moyenne conditionnelle d'une classe de modèles hétérosédastiques non-linéaires.

Depuis Page (1954), la détection de rupture a suscité un intérêt croissant, devenant un sujet central dans les études statistiques et économétriques. Des études ont été faites dans le contexte des données i.i.d. La monographie de Csörgö et Horváth (1997) donne un bon résumé de la théorie dans ce cas. Au cours des récentes années, une attention grandissante est portée sur l'étude des ruptures dans le cas dépendant. Nous pouvons citer entre autres

Andrews (1993), Horváth et al. (2006), Ling (2007) et plus récemment Ngatchou-Wandji et Ltaifa (2023).

Bien qu'il y ait de nombreux travaux sur la détection de rupture, l'analyse des petites ruptures reste relativement peu explorée, sauf peut être lorsqu'elle joue un rôle déterminant dans l'étude du comportement asymptotique de la puissance du test (Andrews (1993)) ou dans l'estimation de la localisation (Bhattacharya (1987), Bai (1994)), notamment lorsque l'amplitude de la rupture tend vers zéro.

D'un point de vue pratique, la détection précoce de petites variations peut s'avérer essentielle, car elles peuvent être les prémices d'événements plus significatifs. Par exemple, dans le domaine médical, l'identification de légers changements dans certains paramètres physiologiques peut signaler l'apparition d'une maladie. De même, en ingénierie, la surveillance des petites fluctuations dans le fonctionnement des machines permet d'anticiper d'éventuelles pannes et d'optimiser la maintenance préventive.

2 Le modèle et les hypothèses

Nous nous intéressons à tester la présence d'une rupture faible dans la moyenne conditionnelle de la classe de modèles CHARN (Härdle et al. (1998)) d'ordre p, $p > 1$, définie par:

$$X_t = T(\rho + \mathbb{1}_{t \geq t_0}\gamma, Z_{t-1}) + V(Z_{t-1})\varepsilon_t, \quad t \in \mathbb{Z}, \tag{1}$$

avec

$$X_t = \begin{cases} Y_{t,1} = T(\rho, Z_{t-1,1}) + V(Z_{t-1,1})\varepsilon_t, & t < t_0, \\ Y_{t,2} = T(\rho + \gamma, Z_{t-1,2}) + V(Z_{t-1,2})\varepsilon_t, & t \geq t_0, \end{cases} \tag{2}$$

où,

- $t_0 \in \mathbb{Z}$, ρ et $\gamma \in \mathbb{R}^d$, $d \geq 1$;
- $T : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ et $V : \mathbb{R}^p \to \mathbb{R}$ des fonctions mesurables et $V(x) > 0$;
- $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc fort de densité f et ε_t est non-corrélée avec $\{X_s\}_{s<t}$;
- Pour $j = 1, 2$, $Z_{t,j} = (Y_{t,j}, \ldots, Y_{t-p+1,j}) \in \mathbb{R}^p$, pour $\ell = 0, 1, \ldots, p-1$, $Z_{t_0-1+\ell} = Z_{t_0-1+\ell,1}$ et $Z_{t_0+\ell} = Z_{t_0+\ell,2}$ avec $Z_t = (X_t, \ldots, X_{t-p+1}) \in \mathbb{R}^p$;
- $(Y_{t,1})$ et $(Y_{t,2})$ sont mutuellement indépendants.

Les hypothèses principales de notre travail sont les suivantes:

(A_1) $\lim\limits_{x\to-\infty} f(x) = \lim\limits_{x\to\infty} f(x) = \lim\limits_{x\to-\infty} f'(x) = \lim\limits_{x\to\infty} f'(x) = 0$;

(A_3) Pour tout $j = 1, 2$, $(Z_{t,j})$ est stationnaire et ergodique;

(A_4) η est une mesure sur $\mathbb{R}^d$, telle que, $\forall r > 0$, si $\mathcal{O}_r$ une boule centrée à l'origine de $\mathbb{R}^d$, alors

$$\int_{\beta \in \mathbb{R}^d \setminus \mathcal{O}_r} |\beta| \, d\eta(\beta) = 0.$$

Soient $X^n := \{X_1, \cdots, X_n\}$, $n \geq 1$, une suite générée par (1) et (2). Nous testons

$$H_0 : \gamma = 0 \text{ contre } \; H_1^{(n)} : \gamma = \frac{\beta_0}{\sqrt{n}} \text{ pour tout } \beta_0 \in \mathbb{R}^d, \; \beta_0 \neq 0. \tag{3}$$

L'hypothèse nulle H_0 correspond à l'absence de rupture, tandis que l'hypothèse alternative $H_1^{(n)}$ suggère l'existence d'une rupture, d'amplitude $\beta_0/\sqrt{n}$ à un moment t_0, dans le modèle défini par (1) et (2). Les paramètres ρ_0 et β_0 ainsi que le point de rupture t_0, sont inconnus.

Nous utilisons le test du RV dont l'expression de la statistique est donnée pour $\rho \in \Theta$, où Θ un compact de $\mathbb{R}^d$, $\beta \in \mathbb{R}_*^d$ et $\tau \in \Pi \subset (0,1)$, par

$$\Lambda_n(\rho, \beta, \tau) = \sum_{t=\lfloor n\tau \rfloor + 1}^{n} \left[l_t\left(\rho + \frac{\beta}{\sqrt{n}}\right) - l_t(\rho) \right], \tag{4}$$

où, $l_t(\rho) = \log f(\varepsilon_t(\rho))$, et

$$\varepsilon_t(\rho) = \frac{X_t - T(\rho, Z_{t-1})}{V(Z_{t-1})}.$$

3 Statistique de test

Dans cette section, notre objectif est de construire une statistique permettant de détecter une petite rupture éventuelle dans le modèle défini par (1) et (2).

Notons, pour tout $\rho \in \Theta$, $\beta \in \mathbb{R}_*^d$ et $\tau \in \Pi$,

$$\Psi_n(\rho, \beta, \tau) = \frac{1}{\sqrt{n}} \sum_{t=\lfloor n\tau \rfloor + 1}^{n} \beta^\top D_t(\rho) \text{ et } v^2(\rho, \beta) = \beta^\top \mathcal{J}(\rho) \beta,$$

où $D_t(\rho)$ désigne la dérivée partielle première par rapport à ρ de l_t,

$$\mathcal{J}(\rho) = I(f) E\left[D_1(\rho) D_1^\top(\rho)\right] \text{ et } I(f) = \mathbb{E}\left[\left(\frac{f'\left[\varepsilon_1(\rho)\right]}{f\left[\varepsilon_1(\rho)\right]}\right)^2\right].$$

Le théorème suivant fournit une approximation de (4), en remplaçant ρ par l'estimateur du maximum de vraisemblance de ρ_0, noté $\widehat{\rho}_{0n}(\tau)$.

Théorème 3.1 *Soit Ω un compact de $\mathbb{R}_*^d$. Si les hypothèses (A_1)-(A_3) sont vérifiées. Alors, sous H_0, on a*

$$\sup_{\beta \in \Omega} \sup_{\tau \in \Pi} \left| \Lambda_n(\widehat{\rho}_{0n}(\tau), \beta, \tau) - \Psi_n(\widehat{\rho}_{0n}(\tau), \beta, \tau) - \frac{1-\tau}{2} \widehat{v}_n^2(\widehat{\rho}_{0n}(\tau), \beta) \right| = o_P(1). \tag{5}$$

où, pour tout $\rho \in \mathbb{R}^d$ et $\beta \in \mathbb{R}_*^d$, $\widehat{v}_n^2(\rho, \beta)$ est un estimateur empirique de $v^2(\rho, \beta)$.

Théorème 3.2 *Si les hypothèses* (A_1)-(A_3) *sont vérifiées. Alors, sous* H_0, *on a*

1. $\Psi_n(\widehat{\rho}_{0n}(\tau_0), \beta_0, \tau_0) \xrightarrow{\mathcal{L}}_n \mathcal{N}\left(0,\ \left(\frac{1}{\tau_0 - 1} v^2(\rho_0, \beta_0)\right)\right)$

2. $\Lambda_n(\widehat{\rho}_{0n}(\tau_0), \beta_0, \tau_0) \xrightarrow{\mathcal{L}}_n \mathcal{N}\left(-\frac{(1-\tau_0)}{2} v^2(\rho_0, \beta_0),\ \left(\frac{1}{\tau_0 - 1} v^2(\rho_0, \beta_0)\right)\right)$.

Remarque 3.3 *Nous déduisons du théorème précédent que, sous* H_0, *les familles* $\left\{H_0^{(n)} = H_0,\ n \geq 1\right\}$ *et* $\left\{H_1^{(n)},\ n \geq 1\right\}$ *sont contiguës.*

D'après les deux théorèmes ci-dessus, nous pouvons utiliser $\Psi_n(\widehat{\rho}_{0n}(\tau), \beta, \tau)$ à la place de $\Lambda_n(\widehat{\rho}_{0n}(\tau), \beta, \tau)$, pour construire la statistique du test. En effet, les deux définissent le même test. Nous considérons alors la statistique

$$\widetilde{R}_n := \int_{\mathbb{R}^d} \sup_{\tau \in \Pi} \left[\widehat{v}_n^{-1}(\widehat{\rho}_{0n}(\tau), \beta) \Psi_n(\widehat{\rho}_{0n}(\tau), \beta, \tau)\right] d\eta(\beta), \tag{6}$$

Le théorème ci-après fournit la distribution asymptotique de $\widetilde{R}_n$ sous H_0 et sous $H_1^{(n)}$.

Théorème 3.4 *Si les hypothèses* (A_1)-(A_4) *sont satisfaites, on a*

1. *Sous* H_0,
$$\widetilde{R}_n \xrightarrow{\mathcal{L}}_n \sup_{\tau \in \Pi} \left(-\frac{B^\circ(\tau)}{\tau}\right).$$

2. *Sous* $H_1^{(n)}$,
$$\widetilde{R}_n \xrightarrow{\mathcal{L}}_n \sup_{\tau \in \Pi} \left[\frac{-B^\circ(\tau)}{\tau} + \left(\frac{\tau_0 \wedge \tau}{\tau} - \tau_0\right) v(\rho_0, \beta_0) \int_{\mathbb{R}^d} v^{-1}(\rho_0, \beta) \beta^\top \mathcal{U}(\rho_0) d\eta(\beta)\right],$$

   où $B^\circ(\tau)$ *un pont brownien sur* Π *et* $\mathcal{U}(\rho_0) = \left(\sqrt{\mathcal{J}_{11}(\rho_0)}, \ldots, \sqrt{\mathcal{J}_{dd}(\rho_0)}\right)^\top$, *les* $\mathcal{J}_{ii}(\rho_0)$ *désignent les éléments diagonaux de la matrice* $\mathcal{J}(\rho_0)$.

Nous rejetons H_0 si $R_n > C_\alpha$, où C_α est le quantile d'ordre $1 - \alpha$ de la loi de $\sup_{\tau \in \Pi} \left(-\frac{B^\circ(\tau)}{\tau}\right)$. Cependant, comme C_α est théoriquement difficile à calculer, nous recommandons de l'approximer par le quantile empirique qui peut être facilement obtenu par simulation numérique.

4 Simulations numériques

Nous présentons les résultats numériques obtenus sur des données simulées générées à partir des modèles suivants:

- Bruit blanc shifté

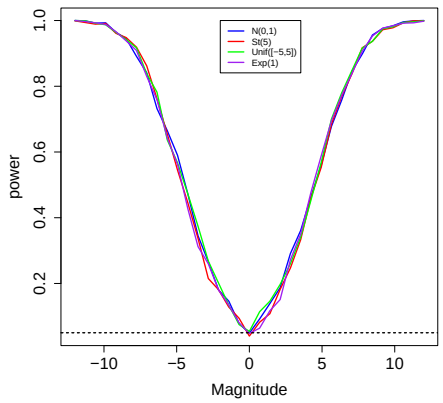
$$X_t = \begin{cases} \rho_0 + \varepsilon_t, \text{ si } t = 1, \dots \lfloor n\tau_0 \rfloor, \\ \rho_0 + \dfrac{\beta_0}{\sqrt{n}} + \varepsilon_t, \text{ si } t = \lfloor n\tau_0 \rfloor + 1, \dots n. \end{cases}$$

- Autorégressive d'ordre 1, AR(1)

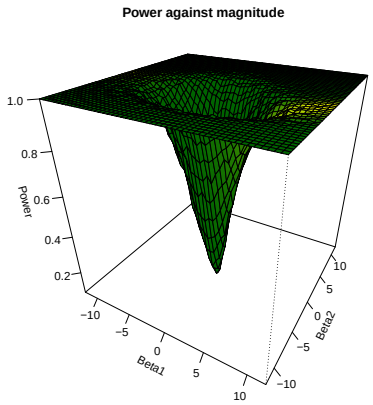
$$X_t = \begin{cases} \rho_1 + \rho_2 X_{t-1} + \varepsilon_t, \text{ si } t = 1, \dots \lfloor n\tau_0 \rfloor, \\ \rho_1 + \dfrac{\beta_1}{\sqrt{n}} + \left(\rho_2 + \dfrac{\beta_2}{\sqrt{n}}\right) X_{t-1} + \varepsilon_t, \text{ si } t = \lfloor n\tau_0 \rfloor + 1, \dots n. \end{cases}$$

(ε_t) est un bruit gaussien centré réduit.

Puis, nous calculons la statistique de test $\widetilde{R}_n$ en utilisant la méthode de Monte Carlo. La valeur obtenue est ensuite comparée à C_α. Ce procédé est répété 1000 fois et la puissance empirique est la proportion de cas dans lesquels $\widetilde{R}_n$ dépasse C_α (la proportion de rejet de l'hypothèse nulle).



(a) Shift gaussien avec $\rho_0 = 3$.



(b) AR(1) avec $\rho_1 = 0.4$ et $\rho_2 = 0.2$.

Figure 1: Puissance empirique de $\widehat{R}_n$ en fonction de β, $n = 100$ et $\tau_0 = 0.4$.

Les deux graphiques montrent que la puissance du test croit en fonction de la norme de β. De plus, la puissance est proche du niveau $\alpha = 0.05$ lorsque β se rapproche de 0.
Le tableau ci-dessous fournit quelques valeurs de la puissance au voisinage de 0:

$\beta = \beta_0 = \beta_1 = \beta_2$	-2	-1	0	1	2
Puissance(figure (a))	0.105	0.097	0.049	0.096	0.106
Puissance(figure (b))	0.119	0.095	0.051	0.092	0.121

Des simulations montrent aussi que pour des valeurs grandes de n et β, la puissance se rapproche de 1 et qu'elle est plus grande pour des ruptures précoces que pour les ruptures tardives.

Bibliographie

Andrews, D. W. K. (1993), Tests for Parameter Instability and Structural Change With Unknown Change Point, *Econometrica*, 61, pp. 821-856.

Bai, J. (1994), Least squares estimation of a shift in linear processes,*Journal of Time Series Analysis*, Wiley Online Library, 15, pp. 453-472.

Bhattacharya, P. K. (1987), Maximum likelihood estimation of a change-point in the distribution of independent random variables: general multiparameter case,*J. Multivariate Anal.*, 23, pp. 183-208.

Bardet, J. and Wintenberger, O. (2009), Asymptotic normality of the quasi-maximum likelihood estimator for multidimensional causal processes, *Ann. Stat*, 37, pp. 2730-2759.

Billingsley, P. (1968), Convergence of probability measures, *Wiley-Blackwell*.

Csörgö, M. and Horváth, L. (1997), Limit theorems in change-point analysis,*Wiley series in probability and mathematical statistics*, John Wiley & Sons, Ltd., Chichester.

Härdle, W. and Tsybakov, A. and Yang, L. (1998), Nonparametric vector autoregression,*Journal of Statistical*, 68, pp. 221-245.

Horváth, L., Kokoszka, P. and Zhang, A. (2006), Monitoring constancy of variance in conditionally heteroskedastic time series, *Econom. Theory*, 22, pp. 373-402.

Ling, S. (2007), Testing for change points in time series models and limiting theorems for NED sequences, *Ann. Stat.*, 35, pp. 1213-1237.

Ngatchou-Wandji, J. and Ltaifa, M. (2023), Detecting weak changes in the mean of a class of nonlinear heteroscedastic models, *Communications in Statistics - Simulation and Computation*, pp. 1-33.

Ngatchou-Wandji, J. and Ltaifa, M. (2024), A Cramér-von Mises test for a class of mean time dependent CHARN models with application to change-point detection, *Statistical Inference for Stochastic Processes*, 27, pp. 25-61.

Page, E. S. (1954). Continuous inspection schemes, *Biometrika*,41, pp. 100-115.

Pollard, D. (1984), Convergence of stochastic processes, *Springer, New York, NY*.

Salman, Y. and Ngatchou-Wandji, J. and Khraibani, Z. (2024), Testing a Class of Piece-Wise CHARN Models with Application to Change-Point Study, *Mathematics*, 12, pp. 2092.

Szyszkowicz, B. (1991) Changepoint problems and contiguous alternatives, *Statistics & probability letters*, 4, pp. 299-308.

Apprentissage par transfert dans la famille exponentielle

Mathias Bourel [1] & Jairo Cugliari [2] & Badih Ghattas [3]

[1] *Universidad de la República, Uruguay,* mbourel@fing.edu.uy
[2] *Lab ERIC, Université Lumière Lyon 2, France,* jairo.cugliari@univ-lyon2.fr
[3] *AMSE, Université d'Aix Marseille,* badih.ghattas@univ-amu.fr

Résumé. L'apprentissage par transfert constitue une méthode clé en apprentissage automatique, permettant de tirer parti des connaissances acquises à partir de modèles pré-entraînés pour résoudre de nouvelles tâches ou améliorer les performances sur des tâches similaires. Ce travail s'intéresse spécifiquement à l'utilisation du transfert d'apprentissage dans le cadre des modèles linéaires généralisés (MLG), en mettant l'accent sur l'approche de réglage fin.

L'objectif principal de cette étude est d'approfondir la compréhension des mécanismes sous-jacents au transfert d'apprentissage dans les MLG. Nous nous concentrons sur une mesure de gain permettant de quantifier le bénéfice (positif ou négatif) d'une approche par transfert. Pour ce faire, nous exploitons les riches liens existants entre les MLG, les distributions de la famille exponentielle et les divergences de Bregman.

Mots-clés. Apprentissage par transfert, modèles linéaires généralisés.

Abstract. Transfer learning is a key method in machine learning, enabling the utilization of knowledge acquired from pre-trained models to solve new tasks or improve performance on similar ones. This work specifically focuses on the application of transfer learning within the framework of generalized linear models (GLMs), with an emphasis on the fine-tuning approach.

The main objective of this study is to deepen the understanding of the underlying mechanisms of transfer learning in GLMs. We focus on a gain measure to quantify the benefit (positive or negative) of a transfer learning approach. To achieve this, we leverage the rich connections between GLMs, exponential family distributions, and Bregman divergences.

Keywords. Transfer learning, generalized linear models

1 Introduction

L'apprentissage par transfert est une technique populaire en apprentissage automatique qui permet d'exploiter les connaissances acquises à partir de modèles pré-entraînés pour résoudre de nouvelles tâches ou améliorer les performances sur des tâches connexes (Pan et Yang, 2009). Il existe plusieurs manières d'appliquer l'apprentissage par transfert, par exemple l'extraction de caractéristiques, le réglage fin (*fine-tuning*), l'adaptation au domaine, l'apprentissage multi-tâches ou encore la distillation des connaissances.

Dans ce travail, nous nous concentrerons sur l'approche de réglage fin afin d'améliorer l'efficacité de l'estimation et prévision d'un modèle linéaire généralisé (MLG). Notre objectif est d'obtenir une compréhension approfondie du transfert dans le cadre des MLG, en explorant les mécanismes sous-jacents et en identifiant les conditions favorisant une généralisation efficace entre différentes tâches.

Un modèle linéaire généralisé repose sur la structure de la famille exponentielle des distributions de probabilité. Trois éléments clés doivent être identifiés pour développer notre cadre : la distribution de la variable de réponse (Section 2), comment conduire le réglage fin des paramètres (Section 3), et l'évaluation du gain (Section 4), afin de garantir une approche rigoureuse et une interprétation fiable des résultats obtenus. La structure de l'article suit chacun de ces éléments. La Section 5 illustre l'approche proposée par des simulations.

2 Modèle Linéaire Généralisé

Un modèle linéaire généralisé repose sur la structure de la famille exponentielle des distributions de probabilité. La densité d'une variable aléatoire Y appartenant à cette famille peut être exprimée sous la forme générale suivante :

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - \Phi(\theta)}{\phi} + c(y,\phi)\right\}, \tag{1}$$

où θ est le paramètre canonique associé à la distribution, ϕ est le paramètre de dispersion (souvent constant dans les MLG classiques), $\Phi(\theta)$ est une fonction (dite fonction cumulante) définissant la relation entre le paramètre canonique et le moment de la distribution, et $c(y,\phi)$ est une fonction qui ne dépend pas de θ. Le principal intérêt de cette formulation réside dans la manière dont les moments de la distribution peuvent être exprimés. En particulier, l'espérance et la variance de Y sont données par $\mathbb{E}[Y] = \mu = \Phi'(\theta)$ et $\mathrm{Var}(Y) = \Phi''(\theta)\phi$. Ces relations permettent de relier directement les paramètres statistiques (les moments de Y) aux paramètres du modèle (θ,ϕ).

Considérons un échantillon $\{(y_i,\mathbf{x}_i) : 1,\ldots,n\}$ d'observations indépendantes issues d'un vecteur aléatoire $(Y,\mathbf{X})$ où Y appartient à une famille exponentielle et $\mathbf{X}$ sont les variables explicatives associées. La fonction de log-vraisemblance pour cet échantillon s'écrit :

$$\ell(\boldsymbol{\theta},\phi;\mathbf{y}) = \sum_{i=1}^{n}\left\{\frac{y_i\theta_i - \Phi(\theta_i)}{\phi} + c(y_i,\phi)\right\}. \tag{2}$$

Ici, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ est le vecteur des observations, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ est le vecteur de paramètres canoniques associés. D'un autre côté une fonction g, appelée fonction de lien, relie $\mu_i = \Phi'(\theta_i)$ à une combinaison linéaire des variables à travers l'expression $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$. Le choix canonique $g = (\Phi')^{-1}$ permet de simplifier les expressions et de rendre ainsi la fonction ℓ dans (2) concave, garantissant une solution unique au problème d'estimation. L'estimation du vecteur des paramètres $\boldsymbol{\beta}$ et éventuellement ϕ, s'effectue par la méthode du maximum de vraisemblance sur la nouvelle fonction de log vraisemblance, $\ell(\boldsymbol{\beta}; \mathbf{y})$, définie de manière analogue à celle de l'équation (2). La maximisation est généralement réalisée à l'aide d'algorithmes itératifs tels que la méthode de Newton-Raphson ou l'algorithme de Fisher scoring. Ces méthodes nécessitent le calcul des dérivées première et seconde de la log-vraisemblance, respectivement :

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i (y_i - \Phi'(\mathbf{x}_i^\top \boldsymbol{\beta})) \qquad \text{et} \qquad \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^2} = -\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \Phi''(\mathbf{x}_i^\top \boldsymbol{\beta}).$$

Pour la suite, il sera utile de définir la matrice diagonale $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$, où $w_i = \Phi''(\mathbf{x}_i^\top \boldsymbol{\beta})$, la matrice $\mathbf{X}$ où la i-ème ligne est $\mathbf{x}_i$ et le vecteur $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^\top$ où $\mu_i = \Phi'(\mathbf{x}_i^\top \boldsymbol{\beta})$. Ainsi, nous obtenons l'expression vectorielle pour le gradient de ℓ ainsi que l'expression matricielle du Hessien, respectivement :

$$\nabla_\ell(\boldsymbol{\beta}) = \mathbf{X}^\top (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})), \qquad \text{et} \qquad \mathbb{H}_\ell(\boldsymbol{\beta}) = -\mathbf{X}^\top \mathbf{W}(\boldsymbol{\beta}) \mathbf{X}.$$

Divergences de Bregman Les divergences de Bregman incluent la plupart des fonctions de perte bien connues, telles que la perte quadratique, la divergence de Kullback-Leibler (KL), la perte logistique, la distance de Mahalanobis, etc. La famille exponentielle et les divergences de Bregman sont profondément connectées (Banerjee et al. 2005). Pour la définir, nous considérons un ensemble convexe $S \subset \mathbb{R}^n$ et une fonction de potentiel $\Phi : S \to \mathbb{R}$ strictement convexe et différentiable sur S. La divergence de Bregman $B_\Phi : S \times S \to [0, +\infty)$ est définie par:

$$B_\Phi(\mathbf{p}|\mathbf{q}) = \Phi(\mathbf{p}) - \Phi(\mathbf{q}) - \langle \nabla_\Phi(\mathbf{q}), \mathbf{p} - \mathbf{q} \rangle.$$

La divergence de Bregman présente certaines caractéristiques d'une distance, notamment la positivité et la séparation. Cependant, elle ne satisfait ni la propriété de symétrie ni celle de l'inégalité triangulaire, ce qui empêche de la considérer comme une véritable distance.

Dans la famille exponentielle, la divergence de Bregman utilisant comme potentiel la fonction cumulant Φ, correspond à la divergence de Kullback-Leibler (étant elle-même aussi une divergence de Bregman) :

$$D_{KL}\left(f_Y(y, \theta) || f_Y(y, \theta')\right) = B_\Phi(\theta'|\theta).$$

La divergence de Bregman peut être interprétée comme une mesure d'écart dans l'espace des paramètres associé à la fonction de potentiel Φ (Eq. (1)) d'un type de distribution de la famille exponentielle. Le paramètre naturel θ et le paramètre moyen $\mu = \nabla\Phi(\theta)$ forment des coordonnées duales dans l'espace des paramètres. Une divergence duale basée sur $\Psi(\mu) = \Phi^*(\mu)$ (la transformée de Legendre de Φ) mesure l'écart dans l'espace des paramètres moyens μ. Une divergence et son dual vérifient l'égalité $B_\Phi(\tilde{\theta}|\theta) = B_\Psi(\mu|\tilde{\mu})$.

3 Transfert par réglage fin

Maintenant nous considérons deux échantillons $\{(y_{\nu,i}, \mathbf{x}_{\nu,i}) : i = 1, \ldots, n_\nu)\}, \nu \in S, C$ que nous appellerons échantillon source (S) et échantillon cible (C) respectivement. Typiquement, le nombre d'observations des sources n_S est sensiblement inférieur au nombre d'observations cible n_C, et la variabilité dans l'échantillon source est relativement plus importante que dans l'échantillon cible. Étant donnée une estimation du paramètre type obtenue à partir du jeux de données source, disons $\hat{\theta}_S = \mathbf{x}^\top \hat{\boldsymbol{\beta}}_S$, nous cherchons à obtenir un estimateur (par transfert) $\hat{\theta}_T = \mathbf{x}^\top \hat{\boldsymbol{\beta}}_T$ préférable (dans un sens à déterminer, voir Section 4) à l'estimateur obtenu à partir des données cible $\hat{\theta}_C = \mathbf{x}^\top \hat{\boldsymbol{\beta}}_C$ pour le paramètre du jeu de données cible $\theta_C = \mathbf{x}^\top \boldsymbol{\beta}_C$.

L'estimateur par transfert $\hat{\theta}_T$ que nous étudions s'obtient par réglage fin de $\hat{\theta}_C$ avec la fonction de perte $J(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n_C} B_\Psi(y_{C,i} | \Phi'(\mathbf{x}_{C,i}^\top \hat{\boldsymbol{\beta}}))$. La solution au problème de minimisation de cette divergence de Bregman coincide avec celle de la maximization de la fonction de vraisemblance car la variable d'intérêt est dans la famille exponentielle de potentiel Φ.

La descente du gradient définit la récurrence suivante : $\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \alpha \nabla_\ell(\hat{\boldsymbol{\beta}}^{(k)})$, où $\alpha > 0$ est le pas du gradient. En utilisant l'expression de $\nabla_\ell(\boldsymbol{\beta})$ et l'identité $\mathbf{X}_C^\top \mathbf{y}_C = \mathbf{X}_C^\top \dot{\Phi}(\mathbf{X}_C \hat{\boldsymbol{\beta}}_C)$, où $\dot{\Phi}(\cdot)$ est la version vectorielle de $\Phi'(\cdot)$, nous obtenons la règle de mis-à-jour :

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \alpha \mathbf{X}_C^\top \left[\dot{\Phi}(\mathbf{X}_C \hat{\boldsymbol{\beta}}_C) - \dot{\Phi}(\mathbf{X}_C \hat{\boldsymbol{\beta}}^{(k)}) \right]. \tag{3}$$

Nous étudions un estimateur alternatif, obtenu comme une approximation linéaire (à chaque itération) de la dernière différence entre les prédicteurs par un point moyen $\boldsymbol{\beta}_\circ^{(k)} = (\tilde{\boldsymbol{\beta}}^{(k)} + \hat{\boldsymbol{\beta}}_C)/2$. Ainsi, l'estimateur d'intérêt est

$$\tilde{\boldsymbol{\beta}}^{(k+1)} = \tilde{\boldsymbol{\beta}}^{(k)} + \alpha \mathbf{X}_C^\top \mathbf{W}(\boldsymbol{\beta}_\circ^{(k)}) \mathbf{X} \left(\hat{\boldsymbol{\beta}}_C - \tilde{\boldsymbol{\beta}}^{(k)} \right). \tag{4}$$

La Proposition 1 rend explicite le fait que l'estimateur par transfert est une combinaison convexe des estimateurs $\hat{\boldsymbol{\beta}}_S$ et $\hat{\boldsymbol{\beta}}_C$ si la matrice $\mathbf{W}$ est constante.

Proposition 1 *Soit* $\mathbf{A} = \mathbf{I} - \alpha \boldsymbol{\Sigma}_C \mathbf{W}$*, alors nous avons pour chaque itération* k *:*

$$\tilde{\boldsymbol{\beta}}^{(k+1)} = \mathbf{A}^{k+1} \hat{\boldsymbol{\beta}}_S + (\mathbf{I} - \mathbf{A}^{k+1}) \hat{\boldsymbol{\beta}}_C.$$

4 Gain

Pour quantifier le bénéfice du transfert, nous définissons une fonction de gain pour une nouvelle observation $(y, \mathbf{x})$ tirée de manière indépendante des jeux de données source et cible, mais identiquement distribuée à ce dernier. Nous comparons les prédictions $\hat{y}_C$ et $\hat{y}_T$ obtenues sans et avec transfert, respectivement. Nous définissons alors le gain de transfert comme une différence entre les pertes de prédiction, c'est-à-dire,

$$\Delta\mathcal{R}(\mathbf{x}) = \texttt{loss}\left(y(\mathbf{x}), \hat{y}_C(\mathbf{x})\right) - \texttt{loss}\left(y(\mathbf{x}), \hat{y}_T(\mathbf{x})\right). \tag{5}$$

Nous souhaitons étudier le gain espéré $\mathcal{E}(\mathbf{x}) = \mathbb{E}[\Delta\mathcal{R}(\mathbf{x})|X = \mathbf{x}]$. La dépendance en $\mathbf{x}$ sera souvent omise dans les notations. La mesure de probabilité qui induit $\mathbb{E}$ est composée de trois sources d'aléa : $S, C, *$, pour le domaine source, le domaine cible et le nouveau domaine de généralisation, respectivement. Ainsi le gain espéré sera étudié comme $\mathcal{E} = \mathbb{E}_{S,C}[\mathbb{E}_*[\Delta\mathcal{R}]]$.

Proposition 2 *Si la fonction de perte est la log-vraisemblance négative, alors, en termes des paramètres naturels $\theta_C = \theta_C(\mathbf{x})$ et des estimateurs $\hat{\theta}_C = \hat{\theta}_C(\mathbf{x}), \hat{\theta}_T = \hat{\theta}_T(\mathbf{x})$, le gain du transfert est une différence entre deux divergences de Bregman, c'est-à-dire :*

$$\mathcal{E} = \mathbb{E}_{S,C}\left[B_\Phi(\hat{\theta}_C|\theta_C) - B_\Phi(\hat{\theta}_T|\theta_C)\right].$$

La Proposition 2 permet de traduire la qualité du problème de prévision dans un problème d'estimation. Le gain est fonction des qualités relatives d'estimation. En intégrant les aléas source et cible (resp. S et C), nous obtenons une décomposition en termes de biais et variance comme l'indique la proposition suivante.

Proposition 3 *Le gain du transfert peut se décomposer en termes des biais et variances généralisées comme :*

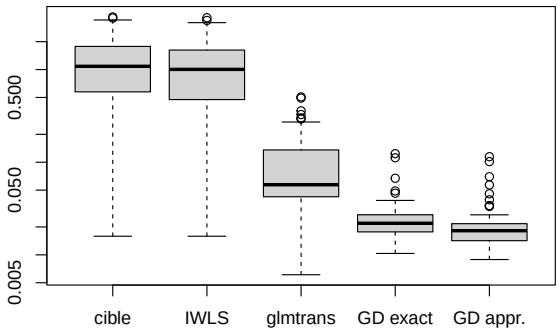
$$\mathcal{E} = B_\Phi(\mathbb{E}(\hat{\theta}_C)|\theta_C) + \mathbb{E}\left[B_\Phi(\hat{\theta}_C|\mathbb{E}(\hat{\theta}_C))\right] - \left\{B_\Phi(\mathbb{E}(\hat{\theta}_T)|\theta_C) + \mathbb{E}\left[B_\Phi(\hat{\theta}_T|\mathbb{E}(\hat{\theta}_T))\right]\right\}.$$

Ainsi, le gain peut être décomposé en deux contributions distinctes, que l'on peut interpréter comme des gains respectifs sur le biais (premier et troisième termes) et sur la variance (deuxième et quatriemme termes). La décomposition permet une meilleure compréhension des mécanismes sous-jacents influençant les performances de l'estimateur considéré.

Dans le cadre spécifique de la régression linéaire gaussienne, le terme biais$_C$ est nul. Par conséquent, le gain ne sera observé que si le terme de biais additionnel pour utiliser l'estimateur par transfert est plus que compensé par la réduction de la variance. En MLG, les estimateurs MV ne sont pas sans biais (seulement asymptotiquement sans biais) ce qui conduit à une plus grande richesse de situations.

5 Illustration numérique

Nous avons testé les approches considérées plus haut par des expériences aléatoires simulées couvrant différents MLG. Nous présentons ici uniquement le cas d'une régression logistique avec deux variables explicatives indépendantes $x_1, x_2 \sim \mathcal{U}(-3, 3)$, $\boldsymbol{\beta}_S = (1, 2, 4)^\top$ et $\boldsymbol{\beta}_C = (1, 3, 6)^\top$. Nous utilisons une approche Monte Carlo (avec 50 répliques), en moyennant ensuite les performances (cf Figure 1). A chaque itération, nous simulons trois échantillons source, cible et test indépendants et de tailles $n_S = 1000$, $n_C = 20$ et $n_* = 1000$ respectivement. Nous estimons les modèles, puis nous mesurons leur qualité en utilisant l'erreur de classification et l'entropie croisée pour les tâches de classification.



Méthode	$B_\Psi(\mu_C\|\cdot)$ moy	sd	0-1 loss (%) moy	sd
cible : $\hat{\boldsymbol{\beta}}_C$	1.282	0.966	7.60	2.96
transf. `IWLS`	1.180	0.898	7.21	2.58
`glmtrans`	0.110	0.120	6.86	2.57
DG approx : $\tilde{\boldsymbol{\beta}}_T$	0.024	0.021	5.25	1.16
DG exact : $\hat{\boldsymbol{\beta}}_T$	0.027	0.021	5.34	1.15
Vrai : $\boldsymbol{\beta}_C$	–	–	4.39	0.65

Figure 1: Moyenne et écart-type des mesures d'erreur pour les variantes utilisées. Résumés obtenus sur 50 répliques pour un modèle de régression logistique en faible dimension.

Pour estimer la cible $y_{C,*}$ nous utilisons : l'approche classique basée sur l'estimation de $\hat{\boldsymbol{\beta}}_C$, l'estimation par IWLS avec initialisation en $\hat{\boldsymbol{\beta}}_S$, l'approche décrite par Tian et Feng (2024) (package `glmtrans`), et les deux variantes étudiées par descente de gradient (DG) exacte ($\hat{\boldsymbol{\beta}}_T$) et approximée ($\tilde{\boldsymbol{\beta}}_T$). Pour faciliter la comparaison, nous incluons la prévision en utilisant le vrai paramètre cible $\boldsymbol{\beta}_C$.

Toutes les méthodes de transfert présentent un gain moyen $\mathcal{E} > 0$ dans cet exemple (pas significatif pour trans. `IWLS`). Les distributions des erreurs utilisant uniquement les données cible ($\hat{\boldsymbol{\beta}}_C$ et transf. `IWLS`) sont relativement plus dispersées, en moyenne élevée que les autres variants. On en déduit que dans cet exemple jouet, l'utilisation des données source est bénéfique. Les méthodes utilisant la descente de gradient affichent ici les gains les plus importants, $\mathcal{E} = 1.261$, tandis que le gain de la méthode basée sur l' `IWLS` est de $\mathcal{E} = 1.182$. Les écarts moyens semblent marginaux, bien que les boxplots mettent en évidence une différence significative entre les dispersions des trois variantes.

6 Conclusion

La sélection de la distribution pour la variable cible est essentielle, car elle doit appartenir à la famille exponentielle pour assurer la cohérence avec le cadre des modèles linéaires généralisés (MLG). Ce choix impacte directement la spécification du modèle et la qualité de l'apprentissage par transfert. Une fois la distribution choisie, la log-vraisemblance négative doit être calculée pour l'échantillon observé. Cette étape est cruciale pour évaluer l'adéquation du modèle et définir la fonction de perte utilisée dans l'optimisation. La descente de gradient permet de résoudre le problème d'optimisation en minimisant la log-vraisemblance négative. La suite des candidats $\hat{\boldsymbol{\beta}}_k$ est générée itérativement selon la règle définie par l'équation (3), garantissant une convergence vers un optimum sous des conditions appropriées. Le choix optimal du nombre d'itérations k est crucial pour équilibrer la convergence et éviter le sur apprentissage. Les prédictions finales sont ensuite calculées en utilisant le paramètre $\boldsymbol{\beta}^{(k)}$ obtenu. L'évaluation du gain de performance est effectuée en calculant la métrique $\mathcal{E}$ définie comme l'espérance de l'expression (5). Cette mesure quantifie l'amélioration apportée par le transfert par rapport à un modèle de référence.

Grâce à ce cadre, nous pouvons traiter simultanément des tâches de classification et de régression, tout en garantissant une flexibilité dans le choix des modèles et une robustesse face aux données hétérogènes, ce qui ouvre des perspectives pour des applications variées dans des contextes pratiques. En perspective, ce travail pourra être étendu à l'étude de modèles non linéaires, permettant une généralisation des résultats obtenus. De plus, le développement d'un test de gain constituerait une avancée significative pour évaluer de manière prédictive le gain du transfert.

Annexes : preuves

►*Preuve de la Prop. 1)* On développe la récurrence dans (3) et comme $\hat{\boldsymbol{\beta}}_S = \hat{\boldsymbol{\beta}}^{(0)}$, on obtient

$$\begin{aligned}\tilde{\boldsymbol{\beta}}^{(k+1)} &= \mathbf{A}^{k+1}\hat{\boldsymbol{\beta}}_S + (\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^k)\hat{\boldsymbol{\beta}}_C \\ &= \mathbf{A}^{k+1}\hat{\boldsymbol{\beta}}_S + (\mathbf{I}-\mathbf{A}^{k+1})\hat{\boldsymbol{\beta}}_C.\end{aligned}$$

►*Preuve de la Prop. 2)* La fonction de perte $\texttt{loss}(y, \hat{y}_C) = -\ell(\theta_C; y)$ pour une v.a suivant (2) s'écrit $\ell(\theta_C; y) = y\theta_C - \Phi(\theta_C)$. Le lien entre $\mathbb{E}(y)$ et Φ donne l'identité suivante :

$$\mathbb{E}_*(\ell(\hat{\theta}_T; y) - \ell(\hat{\theta}_C; y)) = \Phi(\hat{\theta}_C) - \Phi(\hat{\theta}_T) - (\hat{\theta}_T - \hat{\theta}_C)\Phi'(\theta_C).$$

Il suffit de compléter avec $\Phi(\theta_C) - \Phi(\theta_C)$ et $\Phi(\theta_T) - \Phi(\theta_T)$ pour obtenir le résultat.

►*Preuve de la Prop. 3)* On utilise directement la décomposition de Domingos (2000) sur chaque espérance de divergence de Bregman.

Bibliographie

Banerjee, A., Merugu, S., Dhillon, I.S., et Ghosh, J. (2005), Clustering with Bregman Divergences. *Journal of Machine Learning Research*, 6, 1705-1749.

Domingos, P. (2000), A Unified Bias-Variance Decomposition. Proceedings of 17th *International Conference on Machine Learning*, Stanford.

Obst, D., Ghattas, B., Claudel, S., Cugliari, J., Goude, Y., et Oppenheim, G. (2022), Improved Linear Regression Prediction by Transfer Learning. *Computational Statistics & Data Analysis*, 174, 107499. Elsevier.

Pan, S. J., et Yang, Q. (2009), A Survey on Transfer Learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345-1359.

Tian, Y., et Feng, Y. (2023). Transfer Learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544), 2684–2697.

Étude de la médiane empirique pour estimer la moyenne commune de variables aléatoires hétéroscédastiques

Sirine Louati [1]

[1] *CREST-ENSAE Paris, France, sirine.louati@ensae.fr*

Résumé. Nous étudions le problème d'estimation de moyenne dans un cadre hétéroscédastique. Nous considérons pour cela des variables aléatoires symétriques ayant le même paramètre de localisation et des paramètres d'échelle différents et inconnus. Notre objectif est ainsi d'estimer ce paramètre de localisation. Bien qu'élémentaire, ce sujet n'est pas encore bien étudié étant donné que l'on suppose toujours que les variables aléatoires sont indépendantes et identiquement distribuées. Dans ce travail, nous étudions l'estimateur de la médiane empirique et nous établissons des bornes supérieures et inférieures sur son erreur d'estimation qui sont du même ordre de grandeur et qui améliorent de récents résultats de Devroye et al. (2023) et de Xia (2019).

Mots-clés. Estimation optimale; variables aléatoires hétéroscédastiques; médiane empirique.

Abstract. We study the problem of mean estimation in the heteroscedastic setting. In particular, we consider symmetric random variables having the same location parameter and different and unknown scale parameters. Our goal is then to estimate their unknown common location parameter. It is an elementary topic but yet a not very well-studied one since we always make the assumption that the random variables are independent and identically distributed. In this work, we study the median estimator and we establish upper and lower bounds on its estimation error that are of the same order and that generalize and improve recent results of Devroye et al. (2023) and Xia (2019).

Keywords. Optimal estimation; heteroscedastic random variables; empirical median.

1 Introduction

Soient $(X_1, \ldots, X_n)$ des variables aléatoires réelles symétriques ayant le même paramètre de localisation θ et des paramètres d'échelle différents et inconnus σ_i, $i \in [\![1, n]\!]$. Notre objectif est de construire un estimateur de θ qui tienne compte de la structure hétéroscédastique du problème. Ce problème peut être considéré comme un problème "dual" à celui de la robustesse. En effet, dans le cadre de l'estimation robuste, nous cherchons à limiter l'effet néfaste d'un petit nombre de valeurs aberrantes. Dans l'estimation hétéroscédastique, on

souhaiterait plutôt exploiter un faible nombre d'observations très informatives. Si les densités des X_i et donc les paramètres d'échelle σ_i sont connus, le problème devient facilement traitable. L'estimateur optimal dans ce cas est l'estimateur du maximum de vraisemblance, qui est asymptotiquement minimax sous des hypothèses appropriées. Cependant, les paramètres d'échelle sont inconnus et ne peuvent pas être estimés, car une seule observation de σ_i, est disponible pour chaque $i \in [\![1, n]\!]$. Par ailleurs, du fait de la symétrie du problème, l'estimateur de la médiane empirique se présente comme une estimateur naturel à étudier d'autant plus que, contrairement à la moyenne, la médiane n'est pas sensible aux variances élevées. Notre objectif est ainsi de contrôler la probabilité de déviation de la médiane empirique et d'obtenir des bornes supérieures et inférieures sur son erreur d'estimation avec forte probabilité.

2 Analyse de la médiane empirique

2.1 Borne supérieure

Nous énonçons le théorème qui détermine la borne supérieure de l'erreur d'estimation de la médiane empirique.

Théorème 1. *Soient $\delta \in (0, 1)$, $t \geqslant 0$ et C une constante positive telle que $0 < C < 1$ et $\delta > \exp(-2nC^2)$. Soient $(X_1, \ldots, X_n)$ n variables aléatoires indépendantes symétriques autour de la moyenne θ et tel que pour tout $i \in [\![1, n]\!]$ il existe σ_i tel que si $0 \leqslant t \leqslant \sigma_i$ alors*

$$\mathbb{P}(\theta \leqslant X_i \leqslant \theta + t) \geqslant C\frac{t}{\sigma_i}.$$

Alors, on a avec probabilité au moins $1 - \delta$ que,

$$|\widehat{\text{med}}(X_1, \ldots, X_n) - \theta| \leqslant \frac{1}{C\sqrt{2}} \frac{\sqrt{n \log(\frac{2}{\delta})}}{\sum_{i=j+1}^{n} \sigma_i^{-1}},$$

avec $j = \lfloor \frac{1}{C\sqrt{2}} \sqrt{n \log(\frac{1}{\delta})} \rfloor$.

Remarque Il est à noter que la condition du Théorème 1 est satisfaite si $\frac{X_i}{\sigma_i}$ a une densité bornée autour de θ et ainsi ne nécessite pas l'hypothèse de symétrie. Cela signifie que la condition du théorème qui stipule que les variables aléatoires sont symétriques autour de θ et tel que $\mathbb{P}(\theta \leqslant X_i \leqslant \theta + t) \geqslant C\frac{t}{\sigma_i}$ peut être remplacée par la condition suivante

$$\min\left(\mathbb{P}(\theta \leqslant X_i \leqslant \theta + t), \mathbb{P}(\theta - t \leqslant X_i \leqslant \theta)\right) \geqslant C\frac{t}{\sigma_i}.$$

Le corollaire suivant est une application du Théorème 1 dans le cadre gaussien.

Corollaire 1. *Soit* $\delta \in (0,1)$ *tel que* $\delta > \exp\left(-\frac{n}{e\pi}\right)$ *et soient* $(X_1, \ldots, X_n)$ n *variables aléatoires gaussiennes indépendantes tel que* $X_i \sim \mathcal{N}(\theta, \sigma_i^2)$. *Alors, on a avec probabilité au moins* $1-\delta$ *que,*

$$|\widehat{\text{med}}(X_1, \ldots, X_n) - \theta| \leqslant 2.93 \frac{\sqrt{n \log\left(\frac{2}{\delta}\right)}}{\sum_{i=j+1}^{n} \sigma_i^{-1}},$$

avec $j = \lfloor 2.93\sqrt{n \log\left(\frac{1}{\delta}\right)} \rfloor$.

Remarque Ce corollaire découle du Théorème 1 avec $C = \frac{\exp\left(-\frac{1}{2}\right)}{\sqrt{2\pi}} = 0.24$.

2.2 Borne inférieure

Comme indiqué dans l'introduction, aucun estimateur ne peut atteindre une erreur d'estimation d'ordre inférieur à celui de l'estimateur du maximum de vraisemblance, qui a connaissance de toutes les variances et dont la borne est minimax. Cependant, en l'absence de connaissance de ces variances, le problème devient beaucoup plus subtil. Il serait donc intéressant d'identifier une borne inférieure qui soit le plus proche possible de celle de l'estimateur du maximum de vraisemblance.

Nous disposons de la borne inférieure suivante sur l'erreur d'estimation de la médiane empirique.

Théorème 2. *Soit* $\delta \in \left(\exp\left(-(\sqrt{2}-1)^2 n\right), \frac{1}{4}\right)$ *et soient* $(X_1, \ldots, X_n)$ n *variables aléatoires gaussiennes indépendantes tel que* $X_i \sim \mathcal{N}(\theta, \sigma_i^2)$. *Alors, on a avec probabilité au moins* δ *que,*

$$|\widehat{\text{med}}(X_1, \ldots, X_n) - \theta| \geqslant 0.13 \frac{\sqrt{n \log\left(\frac{1}{\delta}\right)}}{\sum_{i=j+1}^{n} \sigma_i^{-1}},$$

avec $j = \lfloor 0.05\sqrt{n \log\left(\frac{1}{\delta}\right)} \rfloor$.

Nous en concluons que la borne inférieure de l'erreur d'estimation de la médiane empirique est bien du même ordre de grandeur que celui de la borne supérieure et ce à une constante près. Par ailleurs, il est à noter que cette borne n'est pas sensible aux $\sqrt{n}$ plus petites valeurs des paramètres d'échelle σ_i. Le comportement de la médiane empirique n'est ainsi pas affecté par ces valeurs.

3 Discussion

Dans le contexte de l'estimation de la moyenne commune de variables aléatoires hétéroscédastiques, plusieurs travaux de recherche ont déjà été effectués, permettant de mieux comprendre et de développer des estimateurs robustes. L'article de Devroye, Lattanzi, Lugosi

et Zhivotovskiy (2023) étudie par exemple l'estimation de la moyenne commune de variables aléatoires indépendantes à densité symétrique et unimodale. Les auteurs proposent un estimateur adaptatif qui combine l'estimateur empirique de la médiane et l'intervalle modal, et démontrent sa robustesse face au bruit et aux valeurs aberrantes.

En outre, les travaux de Xia (2019) mettent en lumière la relation entre la borne de l'erreur d'estimation de la médiane empirique et la moyenne harmonique des écarts-types dans le cadre de l'estimation de la médiane commune de variables aléatoires gaussiennes indépendantes.

La borne supérieure sur l'erreur d'estimation de la médiane empirique que nous avons établie (Louati (2025)) permet d'améliorer et de généraliser les résultats de Devroye et al. (2023) et de Xia (2019). En effet, notre résultat est applicable dans un cadre plus général que celui de Xia (2019) et est plus précis que celui de Devroye et al. (2023).

Bibliographie

Devroye, L., Lattanzi, S., Lugosi, G. et Zhivotovskiy, N. (2023), On mean estimation for heteroscedastic random variables, *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, 59, pp. 1-20.

Louati, S. (2025), The empirical median for estimating the common mean of heteroscedastic random variables, *arXiv preprint arXiv:2501.16956*.

Xia, D. (2019), Non-asymptotic bounds for percentiles of independent non-identical random variables, *Statistics & Probability Letters*, 152, pp. 111-120.

Post hoc bounds for heterogeneous tests

Romain Périer [1] & Gilles Blanchard [2] & Sebastian Döhler [3] & Guillermo Durand [4] & Etienne Roquain [5]

[1] *Université Paris-Saclay, France, romain.perier@universite-paris-saclay.fr*
[2] *Université Paris-Saclay, France, gilles.blanchard@universite-paris-saclay.fr*
[3] *Hochschule Darmstadt, Germany, sebastian.doehler@h-da.de*
[4] *Université Paris-Saclay, France, guillermo.durand@universite-paris-saclay.fr*
[5] *Sorbonne Université, France, etienne.roquain@upmc.fr*

Résumé. Ce travail revisite la construction de bornes post hoc pour les faux positifs dans un contexte de tests multiples, du point de vue des tests hétérogènes, qui apparaissent naturellement dans le cadre de tests discrets. Une grande littérature a récemment émergé autour de la construction de telles bornes, avec notamment la méta-méthode de l'interpolation sur des familles de référence présentée originellement par Blanchard, Neuvial, et al. (2020). En tant que méta-méthode, elle est très flexible et permet de multiples constructions, explorées notamment dans Blanchard, Neuvial, et al. (2020), Durand, Blanchard, et al. (2020) et Meah, Blanchard, et al. (2024). Ces constructions sont cependant conçues pour le "pire cas possible", à savoir des p-valeurs homogènes de distribution uniforme sous l'hypothèse nulle, et lorsque que les distributions des p-valeurs sous la nulle sont hétérogènes super-uniformes, ces constructions peuvent induire une perte de capacité à détecter la présence de signal. Comme, dans ce contexte, les distributions sous la nulle sont connues, nous proposons ici d'exploiter cette information en modifiant certaines constructions des trois articles précédents afin d'incorporer la connaissance de ces distributions et d'atteindre une meilleure puissance que les constructions agnostiques.

Mots-clés. Tests multiples, hétérogénéité, tests discrets, bornes post hoc, inférence sélective

Abstract. This work revisits the construction of post hoc bounds for false positives in a multiple testing context, from the perspective of heterogeneous tests, which naturally arise in the context of discrete tests. A significant body of literature has recently emerged around the construction of such bounds, notably the meta-method of interpolation on reference families originally presented by Blanchard, Neuvial, et al. (2020). As a meta-method, it is very flexible and allows for multiple constructions, notably explored in Blanchard, Neuvial, et al. (2020), Durand, Blanchard, et al. (2020), and Meah, Blanchard, et al. (2024). These constructions are, however, designed for the "worst-case scenario", namely homogeneous p-values uniformly distributed under the null hypothesis, and when the distributions of p-values under the null are heterogeneously super-uniform, these constructions can lead to a loss of ability to detect signal. Since, in this context, the distributions under the null are known, we propose here to exploit this information by modifying some constructions of the three previous papers in order to incorporate the knowledge of these distributions and achieve better power than the agnostic constructions.

Keywords. Multiple testing, heterogeneity, discrete tests, post hoc bounds, selective inference

1 Introduction

The problem of the construction of confidence post hoc bounds for false positives arises from the need to derive statistical guarantees when the experimenter uses the observation of their data (hence the term "post hoc") to select their statistical procedure. Because in doing so, the user preselects data, this problem can be thought of being part of the larger theme of selective inference. First tackled by Genovese and Wasserman (2006) and Meinshausen (2006), it has really been put in the spotlight by Goeman and Solari (2011) in a paper where they advocated for the construction of confidence bounds in the context of exploratory analysis.

Formally speaking, we place ourselves in a multiple testing setting framework where we test m null hypotheses $H_{0,1}, \dots, H_{0,m}$ on a statistical model $(\mathbb{X}, \mathcal{X}, \mathfrak{F})$. We denote by $\mathcal{H}_0 = \mathcal{H}_0(P)$ the indices of the null hypotheses that are true for a probability measure $P \in \mathfrak{F}$ over $\mathcal{X}$. We have at hand m statistics $p_1, \dots, p_m : \mathbb{X} \to [0,1]$ called p-values which are super-uniform under the null:

$$\forall P \in \mathfrak{F}, \forall \alpha \in [0,1], \forall i \in \mathcal{H}_0, P\left(p_i \leq \alpha\right) \leq \alpha.$$

In all the remainder, we consider that the p-values under the corresponding null are independent.

For a given subset of hypotheses $S \subset [\![1, m]\!]$, we are interested in the quantity $V(S) = |S \cap \mathcal{H}_0|$ which is the number of false positives in S. Precisely, we want to construct a (data-dependent) function $\widehat{V}$ such that

$$\forall P \in \mathfrak{F}, \forall \alpha \in]0,1[, P\left(\forall S \subseteq [\![1, m]\!], V(S) \leq \widehat{V}(S)\right) \geq 1 - \alpha. \tag{1}$$

$\widehat{V}$ is called a confidence post hoc bound. Notice that the "$\forall S$" is inside the probability in (1), making it a uniform guarantee and making $\widehat{V}(\widehat{S})$ an over-estimator with high probability of $V(\widehat{S})$ for any selected $\widehat{S}$. Also notice that by dividing each member of the inequality we get a confidence bound over the False Discovery Proportion (FDP):

$$\forall P \in \mathfrak{F}, \forall \alpha \in]0,1[, P\left(\forall S \subseteq [\![1, m]\!], \mathrm{FDP}(S) \leq \frac{\widehat{V}(S)}{|S| \vee 1}\right) \geq 1 - \alpha,$$

hence confidence bounds allow to design sets of hypotheses on which the FDP is controlled with high probability, which is more informative that the classical control holding only on the the expected value of the FDP, the False Discovery Rate (FDR).

We recall the reference family methodology in Section 2.1, we give past constructions of $\mathfrak{R}$ in Section 2.2 and we propose new constructions adapted to heterogeneity in Section 3, after having introduced the problem posed by said heterogeneity.

2 General methodology and previous methods

2.1 General methodology

The construction of post hoc bounds based on reference families (Blanchard, Neuvial, et al., 2020) is the following. A reference family $\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathcal{K}}$ with $R_k \subseteq [\![1, m]\!]$, $\zeta_k \in [\![0, |R_k|]\!]$ (everything can be random) is such that the Joint Error Rate (JER):

$$\mathrm{JER}(\mathfrak{R}) = P\left(\exists k, |R_k \cap \mathcal{H}_0| > \zeta_k\right) \tag{2}$$

is controlled at level α for all $P \in \mathfrak{F}$. Hence $P\left(\forall k, V(R_k) \leq \zeta_k\right) \geq 1-\alpha$ and the ζ_k act as confidence bounds for the number of false positives of the R_k's. The key element here is that $\mathcal{K}$ is generally much smaller than $\mathcal{P}([\![1, m]\!])$, so constructing $\mathfrak{R}$ is tractable. We then deduce a confidence bound over all $\mathcal{P}([\![1, m]\!])$ by interpolation. Namely, with high probability $1-\alpha$, it holds $\mathcal{H}_0 \in \mathcal{A}(\mathfrak{R}) = \{A \subseteq [\![1, m]\!], \forall k, |R_k \cap A| \leq \zeta_k\}$ and therefore we can bound $V(S)$ by

$$V^*_{\mathfrak{R}}(S) = \max_{A\in\mathcal{A}(\mathfrak{R})} |S \cap A|. \tag{3}$$

The maximum in (3) makes for a quantity that is generally intractable, but, with some structure on the sets R_k's, polynomial algorithms to compute $V^*_{\mathfrak{R}}(S)$ exist. One such structure is nestedness, that is the R_k's can be ordered for the set inclusion relation.

2.2 Previous methods

2.2.1 Uniform k-FWER control

This consists in taking $\mathcal{K} = [\![1, m]\!]$, fixing $\zeta_k = k-1$, $R_k = \{i : p_i \leq t_k\}$ and searching for t_k's such that (2) holds. One such choice is $t_k = \alpha\frac{k}{m}$ (Blanchard, Neuvial, et al., 2020), thanks to Simes' inequality (Simes, 1986). In this case the R_k's are nested.

2.2.2 Deterministic regions with true nulls estimation

This consists in assuming that the R_k's are deterministic, given by a priori information. A set structure that is more general than nestedness is assumed, called forest structure, in which for all $k, k' \in \mathcal{K}$, $R_k \cap R_{k'} \in \{R_k, R_{k'}, \varnothing\}$ (Durand, Blanchard, et al., 2020). We then need to compute ζ_k's such that (2) holds, this is done by deriving over-estimators of the $V(R_k)$ from a concentration inequality. The authors of Durand, Blanchard, et al. (2020) propose to use the DKW inequality (Dvoretzky, Kiefer, et al., 1956) with the optimal constant of Massart (1990), which gives the closed form formula:

$$\begin{aligned}\zeta_k &= |R_k| \wedge \min_{t\in[0,1)} \left\lfloor \frac{C}{2(1-t)} + \left(\frac{C^2}{4(1-t)^2} + \frac{\sum_{i\in R_k} \mathbb{1}\{p_i > t\}}{1-t}\right)^{1/2} \right\rfloor^2 \\ &= |R_k| \wedge \min_{0\leq\ell\leq|R_k|} \left\lfloor \frac{C}{2(1-p_{(\ell:R_k)})} + \left(\frac{C^2}{4(1-p_{(\ell:R_k)})^2} + \frac{|R_k|-\ell}{1-p_{(\ell:R_k)}}\right)^{1/2} \right\rfloor^2,\end{aligned}$$

where $C = \sqrt{\frac{1}{2}\log\left(\frac{K}{\alpha}\right)}$, $p_{(0:R_k)} = 0$ and $p_{(\ell:R_k)}$ is the ℓ largest p-value in R_k.

2.2.3 Top-k setting

In this setting, denoting $p_{(\ell)}$ the ℓ largest p-value, we set $R_k = \{i : p_i \leq p_{(k)}\}$ (so they are nested) and once again we over-estimate $V(R_k)$ thanks to a concentration inequality to achieve JER control. The authors of Meah, Blanchard, et al. (2024) review three already existing constructions based on the Simes, DKW, and KR (Katsevich and Ramdas, 2020) inequalities, and introduce a new construction based on a new concentration inequality that is itself derived from Wellner's inequality (Shorack and Wellner, 2009).

3 New methods

We revisit the previous constructions to adapt them to an heterogeneous distribution of the p-values. When the distribution of the p-values is not homogeneous uniform on $[0, 1]$, all previous methods, while remaining valid, incur a loss of power, because they are tailored for the homogeneous uniform case, which can be thought of the "worst case" in multiple testing, with respect to the type I error. Knowing the heterogeneous distributions of the p-values under the null, we can use old and recent tools to derive better reference families in each of the three previous settings.

Let us introduce some notation for the following propositions :

- $\mathcal{A}_i$ is the support of p_i, under $H_{0,i}$, for all $i \in [\![1, m]\!]$,
- $\mathcal{A} = \bigcup_{1\leq i\leq m} \mathcal{A}_i \cup \{0\}$ is the union of all these sets with 0.

In the uniform k-FWER setting, we remark that the values $t_k = \alpha\frac{k}{m}$ are the critical values of the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) and that the Simes inequality can be derived from the FDR control of the BH procedure. Likewise, we propose as t_k the critical values of the HSU or AHSU procedures of Döhler, Durand, et al. (2018). Thanks to the FDR bound proven in the latter paper, we derive a Simes-like inequality which in turn guarantees the JER control of the resulting reference family in the following theorem:

Theorem 3.1. *Let* $\mathbf{H}(s,t) = \left(\frac{F_i(t)}{1-F_i(s)}\right)_{1\leq i\leq m}$ *for all* $(s,t) \in [0,1[\times[0,1]$, $\alpha \in]0,1[$ *and define*
$\tau_m := \max\left\{t \in \mathcal{A} : (\mathbf{H}(t,t))_{(m)} \leq m\alpha\right\}$, *and, for all* $k \in \{1,\dots,m-1\}$,
$\tau_k := \max\left\{t \in \mathcal{A}, t \leq \tau_m : (\mathbf{H}(\tau_m,t))_{(k)} + \cdots + (\mathbf{H}(\tau_m,t))_{(m)} \leq k\alpha\right\}$.
Then, the reference family $\mathfrak{R} = (R_k, \zeta_k)_{1\leq k\leq m}$ *defined by* $R_k = \{i : p_i \leq \tau_k\}$ *and* $\zeta_k = k-1$ *has its* JER *bounded by* α.

Another way to adapt the previous constructions is to use heterogeneous-designed bounds on the empirical cumulative distribution function (ecdf), like Bretagnolle's inequality provided by Shorack and Wellner (2009). This bound is a heterogeneous version of the DKW

inequality and has the advantage of allowing us to compare the ecdf with the mean of p-values cdf under the null hypotheses, denoted by $\overline{F}$. This quantity is, under heterogeneous assumption, strictly upper-bounded by the identity function. This property gives, when the p-values are "strongly heterogeneous", tighter bounds on the number of false positives despite a loosened control on the probability of interest.

In the deterministic setting, we propose a new bound based on Bretagnolle's inequality:

Theorem 3.2. *Let $(R_k)_{k\in\mathcal{K}}$ be a deterministic family and $\alpha \in]0,1[$. Suppose that for all $k \in \mathcal{K}$, $(p_i)_{i\in\mathcal{H}_0\cap R_k}$ is an independent family of p-values. Define $\lambda = \sqrt{\frac{1+\log(2)+\log(K/\alpha)}{2}}$, $\mathbb{G}_k$ such that $\mathbb{G}_k(t) = (\mathbb{1}_{p_i\leq t} - F_i(t))_{i\in R_k}$ for all $t \in [0,1]$, and*

$$b_k(p_i) = \max\left\{\ell : \ell^{-1/2}\left(\sum_{1\leq j\leq \ell}(\mathbb{G}_k(p_i))_{(j)}\right) \leq \lambda\right\}.$$

Then, the reference family $\mathfrak{R} = (R_k, \zeta_k)_{k\in\mathcal{K}}$, with $\zeta_k = \min_{i\in R_k} b_k(p_i)$, has its JER *bounded by α.*

As a concentration inequality, Bretagnolle's inequality can also be used in the top-k setting, which we plan to study.

4 Simulations

4.1 General setting

We simulate a set of 3200 p-values, with $|\mathcal{H}_0| = m_0$ and $|\mathcal{H}_1| = m_1$. Each p-value is defined by $p_i := F_i^{-1}(\tilde{p}_i)$, where each F_i's support are sampled from 19 independent uniform variables with values in $[0,1]$ (substantially inducing power lacking p-values), and each $\tilde{p}_i$ is a p-value sampled from a two-sided independent Gaussian model, where laws of underlying data under $H_{0,i}$ is $\mathcal{N}(0,1)$ and under $H_{1,i}$ is $\mathcal{N}(3,1)$. Despite it is an artificial way to simulate supports, it is really convenient to have fast simulations. We plan to investigate more realistic settings than those exposed here, including applying them to real datasets like in Döhler, Durand, et al. (2018).

For methods based on deterministic regions, we compute our bounds on a forest structure given by a dyadic tree structure, with 8 active leaves and a parameter d corresponding to the proportion of p-values under the alternative in each active leaf. Each leaf of this tree contains 100 hypotheses and the quantity m_1 depends of d by the relation $m_1 = 800d$.

4.2 Significance of the density on Bretagnolle's calibration

In these two simulations, the quantity d will differ. With $d = 1$ in the left panel, Bretagnolle's calibration (bretagnolle.tree curve) is less conservative than DKWM's (DKWM.tree

curve), and the Simes-like inequality derived from Döhler, Durand, et al. (2018) (heterogeneous.simes) at least detects some alternatives in this setting (which is not favourable to KR's and Simes' methods because of the lack of power of our p-values). With a density $d = 0.5$ of p-values in active leaves (right panel), the gain of using Bretagnolle's calibration compared to DKWM's is even greater.
Note that these settings are favourable to Bretagnolle's calibration over DKWM's, and that Bretagnolle's method can be more conservative than DKWM's because of the term $1+\log(2)$ in the definition of λ compared to the constant C of DKWM's method.

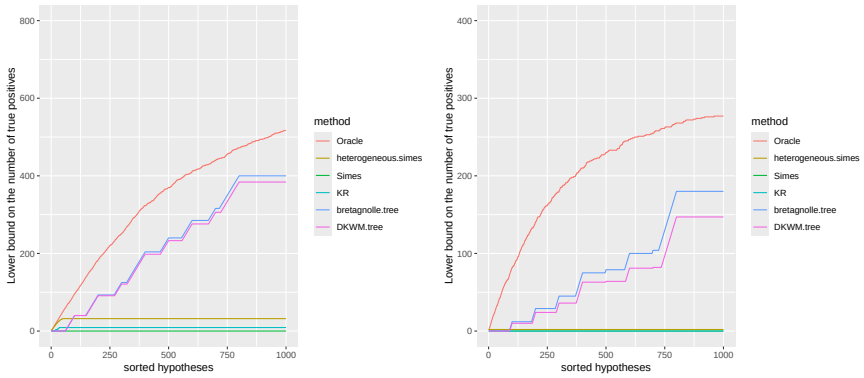


Figure 1: Bounds on top-k sets, computed with $\alpha = 0,05$.

Bibliographie

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *J. Roy. Statist. Soc. Ser. B* 57.1, pp. 289–300. ISSN: 0035-9246. URL: https://www.jstor.org/stable/2346101.

Blanchard, Gilles, Pierre Neuvial, and Etienne Roquain (2020). "Post hoc confidence bounds on false positives using reference families". In: *Ann. Statist.* 48.3, pp. 1281–1303. ISSN: 0090-5364. DOI: 10.1214/19-AOS1847. URL: https://doi.org/10.1214/19-AOS1847.

Döhler, Sebastian, Guillermo Durand, and Etienne Roquain (2018). "New FDR bounds for discrete and heterogeneous tests". In: *Electron. J. Stat.* 12.1, pp. 1867–1900. DOI: 10.1214/18-EJS1441. URL: https://doi.org/10.1214/18-EJS1441.

Durand, Guillermo et al. (2020). "Post hoc false positive control for structured hypotheses". In: *Scand. J. Stat.* 47.4, pp. 1114–1148. ISSN: 0303-6898. DOI: 10.1111/sjos.12453. URL: https://doi.org/10.1111/sjos.12453.

Dvoretzky, A., J. Kiefer, and J. Wolfowitz (1956). "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator". In: *Ann. Math. Statist.* 27, pp. 642–669. ISSN: 0003-4851. DOI: 10.1214/aoms/1177728174. URL: https://doi.org/10.1214/aoms/1177728174.

Genovese, Christopher R. and Larry Wasserman (2006). "Exceedance control of the false discovery proportion". In: *J. Amer. Statist. Assoc.* 101.476, pp. 1408–1417. ISSN: 0162-1459. DOI: 10.1198/016214506000000339. URL: https://doi.org/10.1198/016214506000000339.

Goeman, Jelle J. and Aldo Solari (2011). "Multiple testing for exploratory research". In: *Statist. Sci.* 26.4, pp. 584–597. ISSN: 0883-4237. DOI: 10.1214/11-STS356. URL: https://doi.org/10.1214/11-STS356.

Katsevich, Eugene and Aaditya Ramdas (2020). "Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings". In: *The Annals of Statistics* 48.6, pp. 3465 –3487. DOI: 10.1214/19-AOS1938. URL: https://doi.org/10.1214/19-AOS1938.

Massart, P. (1990). "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality". In: *Ann. Probab.* 18.3, pp. 1269–1283. ISSN: 0091-1798,2168-894X. URL: http://links.jstor.org/sici?sici=0091-1798(199007)18:3<1269:TTCITD>2.0.CO;2-Q&origin=MSN.

Meah, Iqraa, Gilles Blanchard, and Etienne Roquain (2024). "False discovery proportion envelopes with m-consistency". In: *Journal of Machine Learning Research* 25.270, pp. 1–52. URL: http://jmlr.org/papers/v25/23-1025.html.

Meinshausen, Nicolai (2006). "False discovery control for multiple tests of association under general dependence". In: *Scand. J. Statist.* 33.2, pp. 227–237. ISSN: 0303-6898. DOI: 10.1111/j.1467-9469.2005.00488.x. URL: https://doi.org/10.1111/j.1467-9469.2005.00488.x.

Shorack, Galen R and Jon A Wellner (2009). *Empirical processes with applications to statistics*. SIAM.

Simes, R. J. (1986). "An improved Bonferroni procedure for multiple tests of significance". In: *Biometrika* 73.3, pp. 751–754. ISSN: 0006-3444. DOI: 10.1093/biomet/73.3.751. URL: https://doi.org/10.1093/biomet/73.3.751.

Statistique mathématique pour données complexes

PROCESSUS EMPIRIQUES AVEC DONNÉES HÉTÉROGÈNES

Philippe Berthet [1] & Jérémy Boyer [2]

[1] *Institut de mathématiques de Toulouse, France et philippe.berthet@math.univ-toulouse.fr*
[2] *Institut de mathématiques de Toulouse, France et jeremy.boyer@math.univ-toulouse.fr*

Résumé. En statistique mathématique, l'essentiel de la littérature est basé sur des échantillons i.i.d. Ce cadre homogène n'est pas réaliste si la collecte des données est impactée par les circonstances, de la contamination, du renforcement, ou des traitements de type boîte noire. Deux types d'hétérogénéité sont à considérer. Le premier type concerne la dépendance pure, qui permet de s'écarter des lois produits mais préserve les lois marginales. Les extensions du cadre indépendant sont bien connues sous hypothèse de dépendance faible, via des coefficients de mélange qui caractérisent une faible mémoire, dans un cadre stationnaire en loi. Le second type concerne la non stationnarité pure, qui bouge les lois marginales mais préserve l'indépendance. Cette hétérogénéité en loi est plus rarement interrogée, et peu de résultats généraux sont disponibles pour faire face à une dérive de la distribution sous-jacente pendant l'échantillonnage. Des modèles d'évolution probabilistes, par exemple markoviens, véhiculent de la dépendance mais reposent sur une stationnarité asymptotique, alors que des modèles statistiques sous indépendance considèrent des drifts d'espérances, ou des décompositions de type séries temporelles.

Notre direction de recherche explore le cadre strictement non stationnaire. Nous étudions la mesure empirique lorsque les n variables aléatoires échantillonnées sont indépendantes mais que chacune d'entre elles provient d'une distribution différente $\mu_{i/n}$, afin de comprendre quels types d'évolution restent contrôlables. Tout d'abord nous présentons un résultat généralisant Lindeberg-Feller à un k-uplet de fonctions. Puis, sous des hypothèses générales nous obtenons des approximations gaussiennes du processus empirique indexé par une infinité de fonctions, pour tout n : soit d'un point de vue rétrospectif de l'évolution (échantillonnage croissant sur la période d'évolution), soit d'un point de vue séquentiel. Il y a convergence en loi lorsque les processus approximants se stabilisent en un processus limite, qui n'est pas le pont brownien usuel, mais un processus gaussien caractéristique de l'hétérogénéité, lui-même distinct du processus associé au mélange des lois observées. Nous montrons en particulier comment des paramètres liés au transport ou à la variation totale entre les μ_t influent sur la réduction de la vitesse d'approximation forte gaussienne établie par Berthet et Mason dans le cadre i.i.d.

Nous proposons des coefficients d'hétérogénéité permettant de tester la stationnarité sous le contrôle de nos résultats. En guise d'illustration, nous prenons le cas d'un mélange de deux populations avec des coefficients de mélanges qui évoluent avec n, ou celui d'une loi dont les paramètres bougent continûment avec le temps.

Mots-clés. Données hétérogènes, Suite de variables aléatoires non stationnaires, Processus empiriques, Convergence faible, Approximations gaussiennes, Test de stationnarité.

Abstract. In mathematical statistics, most of the literature is based on i.i.d. samples. This homogeneous framework is unrealistic if data collection is impacted by circumstances, contamination, reinforcement or black-box treatments. Two types of heterogeneity need to be considered. The first type concerns pure dependence, which allows deviation from product laws but preserves marginal laws. Extensions of the independent framework are well known under the assumption of weak dependence, via mixing coefficients that characterize weak memory, in a law-stationary framework. The second type concerns pure non-stationarity, which shifts the marginal laws but preserves independence. This heterogeneity in law is more rarely questioned, and few general results are available for dealing with drift in the underlying distribution during sampling. Probabilistic evolution models, e.g. Markovian, convey dependence but rely on asymptotic stationarity, whereas statistical models under independence consider expectation drifts, or time-series-like decompositions.

Our research direction explores the strictly non-stationary framework. We study the empirical measure when the n sampled random variables are independent but each comes from a different $\mu_{i/n}$ distribution, in order to understand which types of evolution remain controllable. First, we present a result generalizing Lindeberg-Feller to a k-tuple of functions. Then, under general assumptions, we obtain Gaussian approximations of the empirical process indexed by an infinite number of functions, for all n: either from an evolutionary retrospective point of view (increasing sampling over the evolutionary period), or from a sequential point of view. Donsker-type convergence occurs when the approximate processes stabilize in a limiting process, which is not the usual Brownian bridge, but a Gaussian process characteristic of heterogeneity, itself distinct from the process associated with the mixture of observed laws. We will show how parameters related to transport or total variation influence the reduction of the approximation rate established by Berthet and Mason in the i.i.d. framework.

We propose heterogeneity coefficients for testing stationarity under the control of our results. By way of illustration, we take the case of a mixture of two populations with mixture coefficients that evolve with n, or that of a law whose parameters continuously move with time.

Keywords. Heterogeneous data, Non-stationary sequence of random variables, Empirical processes, Weak convergence, Gaussian approximations, Stationarity test.

1 Résumé détaillé

1.1 Premier modèle de données hétérogènes

Les données traitées par la statistique inférentielle moderne remettent en question la pertinence du cas théorique de type i.i.d. C'est le cas pour des données online, soumises à des évolution de la loi sous-jacente (contamination, propagation, opinions impactées par des événements etc). Le cadre faiblement dépendant a été bien étudié, à l'aide de coefficients de mélange ou des modèles autorégressifs, alors que le cadre non stationnaire ne fait pas l'objet de développements notables en statistique mathématique, la boîte à outils relevant du domaine méthodologique, dans le cadre des séries temporelles. On peut par exemple noter un article récent de Einmahl et He (2023) en théorie des extrêmes. Cependant les résultats généraux liés à la mesure empirique, de type théorème de Donsker, ne s'appliquent pas dans des cas aussi simples en apparence que des mélanges de populations avec évolution de leur loi et/ou de leur proportion, comme en épidémiologie ou en sondage.

Nous proposons d'aborder la question des données hétérogènes dans le cadre général suivant. Soit $X_1, ..., X_n$ des variables aléatoires (v.a.) indépendantes sur un espace mesuré $(\mathcal{X}, \mathcal{A})$ mais pas nécessairement de même loi, et leur mesure empirique

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \quad P_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

indexée par des fonctions réelles $f \in \mathcal{F}$, avec $\mathcal{F}$ uniformément bornée par $M > 0$. Nous représentons l'évolution en temps continu de la loi des observations par un ensemble de mesures de probabilité $\mu = (\mu_t)_{t\in[0,1]}$. Cela permet d'énoncer la convergence faible dans un cadre asymptotique clair, en considérant un tableau triangulaire $(X_n)_{n\geq 1}$ tel qu'à l'étape n,

$(\mathcal{M}_1)$ $\quad X_n = (X_{i,n})_{1\leq i\leq n}$ est constituée de v.a. indépendantes, de loi $\mathcal{L}(X_{i,n}) = \mu_{\frac{i}{n}}$.

Nos résultats portent sur le processus empirique

$$\Gamma_n(f) = \left(\sqrt{n}\left(P_n(f) - \mu_n(f)\right)\right)_{f\in\mathcal{F}}, \quad \mu_n = \frac{1}{n}\sum_{i=1}^{n} \mu_{\frac{i}{n}}.$$

Ils ne peuvent se déduire des convergences classiques i.i.d. répertoriés par exemple dans Van der Vaart et Wellner (1996), dont ils constituent une extension au cadre non-stationnaire. Nous réalisons pour tout n un couplage du processus empirique par des processus gaussiens locaux, de type Berthet et Mason (2006), en utilisant des extensions ad hoc des résultats sur les processus indexés par $\mathcal{F}$ comme Talagrand (1994), ou Borel (1975).

L'expérience statistique associée au modèle $(\mathcal{M}_1)$ est l'apprentissage de l'ensemble $(\mu_t)_{t\in[0,1]}$ pour lequel on dispose d'un échantillonnage sur $[0,1]$ qui est de plus en plus riche mais régulier. Cela permet de mettre en relation la fréquence d'échantillonnage, uniforme, et la vitesse et forme d'évolution des lois. Il ne s'agit pas d'un échantillonnage sous la loi mélange $\overline{\mu} = \int_0^1 \mu_t dt$ mais d'observations ordonnées dans le temps avec une loi déterminée à chaque instant : c'est pourquoi le processus limite de Γ_n n'est pas un pont Brownien indexé par $\overline{\mu}$.

1.2 Un premier théorème central limite

Le théorème de Lindeberg et Feller est un TCL univarié et auto-normalisé pour des sommes de v.a. réelles indépendantes. En étendant la méthode de preuve, présentée dans le livre de Petrov (1975) par exemple, nous obtenons un premier TCL multivarié adapté au modèle $(\mathcal{M}_1)$. On considère le module de continuité de $\mu = (\mu_t)_{t\in[0,1]}$ en variation totale par :

$$\omega_{VT}(\mu, \delta) = \sup \{d_{VT}(\mu_t, \mu_s) \, : \, |t - s| \leq \delta\} .$$

Proposition 1. *Si* $(\mathcal{M}_1)$ *vérifie* $\lim_{\delta\to 0} \omega_{VT}(\delta) = 0$ *alors, pour* $(f_1, ..., f_d) \in \mathcal{F}^d$,

$$(\Gamma_n(f_i))_{1\leq i\leq d} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_d(0, \Sigma), \quad \Sigma = \left(\int_0^1 (\mu_t(f_i f_j) - \mu_t(f_i)\mu_t(f_j))dt \right)_{1\leq i,j\leq d} .$$

L'hypothèse de continuité en variation totale impose à notre modèle d'avoir une évolution régulière, sans rupture forte entre deux observations proches dans le temps.

Ce résultat renseigne sur la structure de covariance du modèle non stationnaire, néanmoins la méthode échoue à étudier la mesure empirique P_n sur une classe infinie de fonctions. Pour répondre à cela, on va généraliser le principe d'approximation forte suivant.

1.3 Approximation du processus empirique stationnaire

Supposons la loi sous-jacente μ_0 stationnaire et rappelons que le μ_0-pont Brownien indexé par $\mathcal{F} \subset L^2(\mu_0)$ est le processus gaussien $\mathbb{G}^{\mu_0}$ centré, de covariance $\mu_0(f_i f_j) - \mu_0(f_i)\mu_0(f_j)$. Dans ce cadre i.i.d. la suite de processus empiriques $\Gamma_n^0 = \sqrt{n}(P_n - \mu_0)$ peut être couplée à une suite de μ-pont browniens, sous des hypothèses d'entropie métrique de $\mathcal{F}$. Soit $d_{\mu_0}(f, g) = ||f - g||_{L^2(\mu_0)}$.

Définition 2. On définit l'entropie uniforme par $N(\varepsilon, \mathcal{F}) = \sup_{\mu_0} N(\varepsilon M, \mathcal{F}, d_{\mu_0})$ et l'entropie crochet par $N_{[\,]}(\varepsilon, \mathcal{F}, d_{\mu_0})$ où $N(\varepsilon, \mathcal{F}, d_{\mu_0})$ et $N_{[\,]}(\varepsilon, \mathcal{F}, d_{\mu_0})$ sont le nombre minimal de boules (respectivement de crochets) de d_{μ_0}-rayon ε nécessaire pour recouvrir $\mathcal{F}$.

Proposition 3. *(Berthet et Mason (2006)) Pour tout* $\lambda > 0$ *il existe* (c_λ, N_λ) *tels qu'on peut construire sur le même espace de probabilité la suite* (X_n) *et une suite* $(\mathbb{G}_n^{\mu_0})$ *de versions de* $\mathbb{G}^{\mu_0}$ *telles que*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} ||\Gamma_n^0 - \mathbb{G}_n^{\mu_0}||_{\mathcal{F}} > c_\lambda v_n \right) < \frac{1}{n^\lambda}, \quad n \geq N_\lambda$$

avec

(VC) si $N(\varepsilon, \mathcal{F}) \leq c\varepsilon^{-v}$, $(c, v > 0)$ *alors* $v_n = (\log n)^\tau n^{-a}$ *avec* $a = \dfrac{1}{2 + 5v}$, $\tau = \dfrac{1}{2} + a$ *(Classe-VC, polynomiale, petites classes),*

(BR) si $\log N_{[\,]}(\varepsilon, \mathcal{F}, d_{\mu_0}) \leq c\varepsilon^{-v}$, $(c > 0, v < 2)$ *alors* $v_n = (\log n)^{-r}$ *avec* $r = \dfrac{1}{v}\left(1 - \dfrac{v}{2}\right)$ *(Classe-BR, exponentielle, grosses classes).*

Parmi les conséquences, ce résultat donne une vitesse de convergence faible v_n de Γ_n^0 vers $\mathbb{G}^{\mu_0}$ pour une classe P-Donsker $\mathcal{F}$. Cela permet aussi d'analyser pour tout n des statistiques de type plug-in $\varphi(P_n) \sim \varphi(P + \mathbb{G}_n^{\mu_0}/\sqrt{n}) \sim \varphi(P) + \nabla\varphi(\mathbb{G}_n^{\mu_0})/\sqrt{n}$ en cas de différentiabilité locale. Nous présentons nos extensions au cadre non-stationnaire $(\mathcal{M}_1)$.

1.4 Approximations du processus empirique non-stationnaire

Représentons $(\mathcal{M}_1)$ à l'aide d'opérateurs de transports ρ_t tels que $\mu_t = \mu_0 \circ \rho_t$ et

$$\lim_{\delta \to 0} \sup_{(f,g)\in\mathcal{F}^2} \sup_{|t-s|<\delta} \mu_0 \left(|fg \circ \rho_t - fg \circ \rho_s|\right) = 0.$$

Ainsi le tableau triangulaire $(\mathcal{M}_1)$ s'écrit à l'étape n sous la forme $\rho_{i/n}(X_i)$ avec $(X_n)_{n\geq 1}$ i.i.d. de loi μ_0. L'évolution des lois est représentée par le flot de fonction $\rho = (\rho_t)_{t\in[0,1]}$ avec ρ_0 l'identité. Notez que ρ n'est pas unique pour représenter $\mu = (\mu_t)_{t\in[0,1]}$ mais reste contraint par le module de continuité ci-dessus, qui garantit la stabilisation de la loi limite des processus gaussiens approximants $\mathbb{G}_n^\rho$ des théorèmes suivants, qui sont de même loi que $\sum_{i=1}^n \mathbb{G}^{\mu_{i/n}}/\sqrt{n}$. La Proposition 4 et les Théorèmes 5 et 6 impliquent $\Gamma_n \xrightarrow{loi} \mathbb{G}_\infty^\mu$.

Proposition 4. *Sous (VC) on a* $\mu_n \xrightarrow{loi} \overline{\mu} = \int_0^1 \mu_t dt$ *et* $\mathbb{G}_n^\rho \xrightarrow{loi} \mathbb{G}_\infty^\mu$ *dans* $l^\infty(\mathcal{F})$ *où* $\mathbb{G}_\infty^\mu$ *est le processus gaussien centré de covariance* $Cov(\mathbb{G}_\infty^\mu(f), \mathbb{G}_\infty^\mu(g)) = \int_0^1 (\mu_t(fg) - \mu_t(f)\mu_t(g))dt$.

Renforçons le contrôle du flot de mesures μ avec

$$(VC/\mu) \qquad \sup_{(f,g)\in\mathcal{F}^2} \sup_{|t-s|<\delta} \left| ||f-g||_{L^2(\mu_t)} - ||f-g||_{L^2(\mu_s)} \right| < \delta^b, \ b > 1/2.$$

Théorème 5. *Sous (VC) et (VC/*μ*) alors pour tout* $\lambda > 0$ *il existe* (c_λ, N_λ) *tels qu'on peut construire sur le même espace de probabilité la suite* $((X_{i,n})_{1\leq i\leq n})$ *et une suite de processus gaussiens centrés* $(\mathbb{G}_n^\rho)$ *tels que*

$$\mathbb{P}\left(||\Gamma_n - \mathbb{G}_n^\rho||_{\mathcal{F}} > c_\lambda v_n\right) < \frac{1}{n^\lambda}, \quad n \geq N_\lambda,$$

avec $v_n = (\log n)^\tau n^{-v(a,b)}$, $v(a,b) = a - \dfrac{a}{a+b}\left(a + \dfrac{1}{2}\right) > 0$.

Renforçons à présent le contrôle du flot de transports ρ par

$$(BR/\mu) \qquad \sup_{f\in\mathcal{F}} \sup_{|t-s|<\delta} ||f \circ \rho_t - f \circ \rho_s||_\infty < \delta^b, \ b > 1/2.$$

Théorème 6. *Si* $\sup_{t\in[0,1]} \log N_{[\,]}(\varepsilon, \mathcal{F}, d_{\mu_t}) \leq c\varepsilon^{-v}$, *(BR/*$\mu$*) et* $v < (2b-1)/b$ *alors pour tout* $\lambda > 0$ *il existe* (c_λ, N_λ) *tels qu'on peut construire sur le même espace de probabilité la suite* $((X_{i,n})_{1\leq i\leq n})$ *et une suite de processus gaussiens centrés* $(\mathbb{G}_n^\rho)$ *tels que*

$$\mathbb{P}\left(||\Gamma_n - \mathbb{G}_n^\rho||_{\mathcal{F}} > c_\lambda v_n\right) < \frac{1}{n^\lambda}, \quad n \geq N_\lambda,$$

avec $v_n = (\log n)^{-w(v,b)}$, $w(v,b) = \dfrac{1}{v}\left(1 - \dfrac{v}{2}\right) - \dfrac{1}{2vb} > 0$.

1.5 Un exemple : mélange évolutif de deux lois.

Etudions le modèle $\mu_t = (1-\alpha(t))\mu_0 + \alpha(t)\mu_1$ où α est une fonction b-höldérienne sur $[0,1]$ avec $b > 1/2$, $\alpha(0) = 0$ et $\alpha(1) = 1$. Si $\mathcal{X} = \mathbb{R}$ il s'agit d'un modèle de transport ρ explicite. A l'aide d'un échantillonnage régulier $X_1, ..., X_n$ tel que $\mathcal{L}(X_i) = \mu_{i/n}$ nous souhaitons tester l'existence d'une telle fonction α puis estimer μ_0, μ_1 et l'évolution $t \mapsto \alpha(t)$.

Soit $k_n \in \mathbb{N}^*$, $l_n = [n/k_n]$, $A_j = [\![(j-1)l_n+1, jl_n]\!]$, $t_j = jl_n/n$ et $P_n^j(f) = l_n^{-1}\sum_{i\in A_j} f(X_i)$. Pour chaque $f \in \mathcal{F}_n = \{f \in \mathcal{F} : |P_n^{k_n}(f) - P_n^1(f)| > c_0\}$ on définit l'estimateur

$$\overline{\alpha}_f(t) = \frac{P_n^j(f) - P_n^1(f)}{P_n^{k_n}(f) - P_n^1(f)},\ t \in]t_{j-1}, t_j].$$

Si l'hypothèse de flot de mélange est vérifiée, sans connaître (μ_0, μ_1, α) les fonctions $\overline{\alpha}_f(t)$ varient peu avec f. Pour une précision uniforme en f et t, appliquons les Théorèmes 5 et 6.

Proposition 7. *Si on a $k_n \sim n^\beta$ et un flot de mélange $\mu_t = (1-\alpha(t))\mu_0 + \alpha(t)\mu_1$ alors on peut construire $((X_{i,n})_{1\le i\le n})$ et une suite de processus gaussiens centrés $(\mathbb{G}_n^\alpha)$ explicites tels que*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}_n}\sup_{t\in[0,t_{k_n}]} |\sqrt{l_n}(\overline{\alpha}_f(t) - \alpha(t)) - \mathbb{G}_n^\alpha(t,f)| > c_\lambda v_n\right) < \frac{1}{n^\lambda}, \quad n \ge N_\lambda,$$

où $v_n \to 0$ dépend de β et de l'entropie métrique de $\mathcal{F}$. De plus, $\mathbb{G}_n^\alpha \xrightarrow{loi} \mathbb{G}_\infty^\alpha$.

D'où un test de l'hypothèse de flot de mélange, avec la statistique de test

$$T_n = \sup_{(f,g)\in\mathcal{F}_n^2}\sup_{t\in[0,t_{k_n}]} \sqrt{l_n}|\overline{\alpha}_f(t) - \overline{\alpha}_g(t)| \xrightarrow{loi} T_\infty = \sup_{(f,g)\in\mathcal{F}_c^2}\sup_{t\in[0,1]} |\mathbb{G}_\infty^\alpha(t,f) - \mathbb{G}_\infty^\alpha(t,g)|.$$

1.6 Second modèle de données hétérogènes

Considérons le cas d'un échantillonnage en rafale à des instants réguliers j/k_n de manière à former $k_n = \lfloor n^\beta \rfloor$ blocs de données, chacun i.i.d. Le tableau triangulaire devient

$(\mathcal{M}_2)$ $X_n = (X_{i,n})_{1\le i\le n}$ v.a. indépendantes, de loi $\mathcal{L}(X_{i,n}) = \mu_{j/k_n}$ pour tout i du bloc j.

Représentons la taille des blocs par une suite de fonctions $(\gamma_n)_{n\ge1} \in C^0([0,1])^{\mathbb{N}^*}$ telle que

$$(S) \quad ||\gamma_n - \gamma||_\infty \longrightarrow 0, \quad \text{et le bloc } j \text{ est de taille } n\gamma_n(j/k_n)/k_n.$$

Pour $\mu_n^\gamma = k_n^{-1}\sum_{j=1}^{k_n}\gamma_n(j/k_n)\mu_{j/k_n}$ et $\Gamma_n^\gamma(f) = n^{-1/2}\sum_{i=1}^n (f(X_i) - \mu_n^\gamma(f))$ on obtient le même type de stabilisation que pour le modèle $(\mathcal{M}_1)$.

Proposition 8. *Si $(\mathcal{M}_2)$ vérifie (VC), (S) et $\lim_{\delta\to0}\omega_{VT}(\mu,\delta) = 0$ alors*

$$\mu_n^\gamma \xrightarrow{loi} \int_0^1 \mu_t\gamma(t)dt, \quad \mathbb{G}_n^\gamma = \frac{1}{k_n}\sum_{j=1}^{k_n}\sqrt{\gamma_n(j/k_n)}\mathbb{G}_n^{\mu_{j/k_n}} \xrightarrow{loi} \mathbb{G}_\infty^\gamma \text{ dans } l^\infty(\mathcal{F}),$$

avec $\mathbb{G}_\infty^\gamma$ gaussien, centré de covariance $Cov(\mathbb{G}_\infty^\gamma(f), \mathbb{G}_\infty^\gamma(g)) = \int_0^1(\mu_t(fg) - \mu_t(f)\mu_t(g))\gamma(t)dt$.

Théorème 9. *Sous (VC), (S) et $\beta < a/(a+1/2)$ pour tout $\lambda > 0$ il existe (c_λ, N_λ) tels qu'on peut construire sur le même espace de probabilité les suites $((X_{i,n})_{1\leq i\leq n})$ et $(\mathbb{G}_n^\gamma)$ vérifiant*

$$\mathbb{P}\left(||\Gamma_n^\gamma - \mathbb{G}_n^\gamma||_{\mathcal{F}} > c_\lambda v_n\right) < \frac{1}{n^\lambda}, \quad n \geq N_\lambda,$$

avec $v_n = (\log n)^\tau n^{-u(a,\beta)}$, $u(a,\beta) = a - \beta\,(a+1/2) > 0$.

2 Bibliographie

Berthet P. and Mason D.M. (2006), Revisiting two strong approximation results of Dudley and Philipp, *IMS Lecture Notes–Monograph Series High Dimensional Probability*, pp. 155-172.

Van der Vaart A.W. and Wellner J.A. (1996), *Weak convergence and empirical processes*, Springer.

Einmahl, J.H.J. and He, Y. (2023), Extreme value inference for heterogeneous power law data., *The Annals of Statistics*, 51 pp. 1331-1356.

Petrov, V. V. (1975), *Sums of Independent Random Variables.*, Springer.

Borell C. (1975), The Brunn-Minkowski inequality in Gauss space., *Inventiones mathematicae*, 51 pp. 207-216.

Talagrand M. (1994), Sharper bounds for Gaussian and empirical processes., *The annals of Probability.*

Tests d'uniformité sur des espaces polonais homogènes compacts

Claire Brécheteau [1]

[1] *Centrale Nantes, Nantes Université, Laboratoire de Mathématiques Jean Leray, France, claire.brecheteau@ec-nantes.fr*

Résumé. Nous considérons un espace polonais compact $(\mathscr{X}, \mathrm{d})$, homogène, au sens où il peut être doté d'une mesure de probabilité uniforme μ_0 : une mesure qui attribue la même masse à toutes les boules partageant le même rayon, pour tout rayon. Étant donné un échantillon de points sur $\mathscr{X}$, nous introduisons et étudions deux familles différentes de tests statistiques basés sur les distances. Les premiers testent que l'échantillon est uniformément réparti sur $\mathscr{X}$, au moins aussi bien qu'un échantillon indépendamment identiquement distribué (i.i.d.) de loi μ_0. Les seconds testent que l'échantillon est i.i.d. de loi μ_0. Les statistiques de test sont basées sur des signatures-DTM, telles qu'introduites dans Brécheteau (2019) pour des tests d'isomorphisme entre espaces métriques mesurés, basés sur deux échantillons. Ces tests sont consistants. Nous illustrons leurs performances sur différents exemples incluant des variétés riemaniennes compactes homogènes sans bord : le cercle, la sphère, la Grassmannienne.

Mots-clés. Tests statistiques, uniformité, distance de Wasserstein

Abstract. We consider a compact Polish space $(\mathscr{X}, \mathrm{d})$ that is homogeneous in the sense that it can be equipped with a uniform probability measure μ_0: a measure that puts the same mass to all balls sharing the same radius, for every radius. Given a sample of points on $\mathscr{X}$, we introduce and study two different distance-based families of statistical tests. The first ones test that the sample is uniformly spread on $\mathscr{X}$, at least as well as an independently identically distributed (i.i.d.) sample from μ_0, the second ones test that the sample is i.i.d. from μ_0. The tests statistics are based on empirical distance-to-measure signatures, as introduced in a paper by Brécheteau (2019) for two-sample tests of isomorphism between metric measure spaces. These tests are consistent. We illustrate their performance on different examples including compact homogeneous Riemanian manifolds with no boundary: the circle, the sphere, the Grassmannian.

Keywords. Statistical tests, uniformity, Wasserstein distance

1 Introduction

1.1 Espaces homogènes compacts et premiers exemples

La notion de mesure de probabilité *uniforme* sur un espace polonais compact $(\mathscr{X}, \mathrm{d})$ apparaît dès Loomis (1945). Il s'agit d'une mesure de probabilité borélienne $\mu_0 \in \mathscr{P}(\mathscr{X})$ qui satisfait

$$\forall x, y \in \mathscr{X},\ \forall \epsilon > 0,\ \mu_0(\mathrm{B}(x, \epsilon)) = \mu_0(\mathrm{B}(y, \epsilon)). \tag{1}$$

L'espace $(\mathscr{X}, \mathrm{d})$ est dit *homogène* si une mesure de probabilité uniforme μ_0 existe.

D'après un résultat de Christensen (1970), pour un espace métrique compact $(\mathscr{X}, \mathrm{d})$, une mesure μ_0, si elle existe, est unique.

Des exemples d'espaces homogènes compacts classiques sont : le tore, le cercle, la sphère, la Grassmanienne ou encore la surface de Bolza, dans le disque de Poincaré. De nombreux tests statistiques d'uniformité existent sur le cercle et la sphère par exemple, avec une synthèse dans García-Portugués et al.(2018). Il en existe aussi sur la Grassmanienne, voir Chikuse (2003), mais pas sur la surface de Bolza, à notre connaissance.

1.2 Se comparer à la mesure uniforme

La distance de Wasserstein permet de comparer deux mesures de probabilité, elle métrise la convergence en loi. Pour $p \in [1, \infty)$, la distance de Wasserstein d'ordre p, $\mathcal{W}_{p,\mathrm{d}}$ (ou encore $\mathcal{W}_p$) est définie pour toutes mesures $\mu, \nu \in \mathscr{P}(\mathscr{X})$ par

$$\mathcal{W}_{p,\mathrm{d}}^p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathscr{X}\times\mathscr{X}} \mathrm{d}^p(x, y) \mathrm{d}\pi(x, y), \tag{2}$$

où $\Pi(\mu, \nu)$ est l'ensemble des mesures de probabilité $\pi \in \mathscr{P}(\mathscr{X} \times \mathscr{X})$ de première marginale μ et de seconde marginale ν.

Afin de tester qu'un échantillon $X_1, \ldots, X_n$ a été généré comme un n-échantillon de points indépendants d'une loi μ donnée, Hallin et al.(2021) compare la loi empirique associée à l'échantillon, $\boldsymbol{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, à la loi μ, à l'aide de $\mathcal{W}_p$. Contrairement à la méthode d'Hallin et al.(2021) nous utilisons des signatures-DTM, pour tester qu'un échantillon est iid de loi μ_0. Ces signatures ont le mérite d'être plus rapides à calculer (calcul de distances aux plus proches voisins). Les signatures-DTM sont des mesures de probabilité sur $\mathbb{R}$, $s_h(\mu)$ (resp. $s_h(\boldsymbol{\mu}_n)$) associées à une mesure μ (resp. à un échantillon de mesure empirique $\boldsymbol{\mu}_n$), et qui ont la propriété d'être stables par rapport à la distance de Wasserstein. Par ailleurs, on peut montrer que ces signatures discriminent la loi uniforme au sens où

$$\forall h \in [0, 1], s_h(\mu) = s_h(\mu_0) \Leftrightarrow \mu = \mu_0. \tag{3}$$

Les signatures-DTM sont définies dans Brécheteau (2019) par $s_h(\mu) = \mathrm{d}_{\mu,h|\#\mu}$, pour un paramètre $h \in [0, 1]$, où $\mathrm{d}_{\mu,h}$ est la fonction distance-à-la-mesure (DTM) définie dans Chazal et al.(2010) et $\mathrm{d}_{\mu,h|\#\mu}$ est la mesure image de μ par $\mathrm{d}_{\mu,h}$. La signature-DTM de paramètre h est donc la loi de $\mathrm{d}_{\mu,h}(X)$, pour X de loi μ. La DTM est définie par $\mathrm{d}_{\mu,h} : x \in \mathscr{X} \mapsto \mathcal{W}_2(\delta_x, \mathrm{Sub}_h(\mu)) \in \mathbb{R}_+$, où $\mathrm{Sub}_h(\mu) = \{\nu \in \mathscr{P}(\mathscr{X}),\ h\nu \leq \mu\}$ est l'ensemble des sous-mesures de μ de masse h. Nous retiendrons que

$$s_h\left(\frac{1}{n}\sum_{j=1}^n \delta_{X_j}\right) = \frac{1}{n}\sum_{j=1}^n \delta_{\sqrt{\frac{1}{nh}\sum_{i=1}^{nh} \mathrm{d}^2(X^{(i)}(X_j), X_j)}} \in \mathscr{P}(\mathbb{R}) \tag{4}$$

et où $X^{(i)}(x)$ est le i-ème plus proche voisin de x dans $\{X_1, \ldots, X_n\}$. Il s'agit de la mesure empirique des distances moyennes des points de l'échantillon à leurs nh plus proches voisins

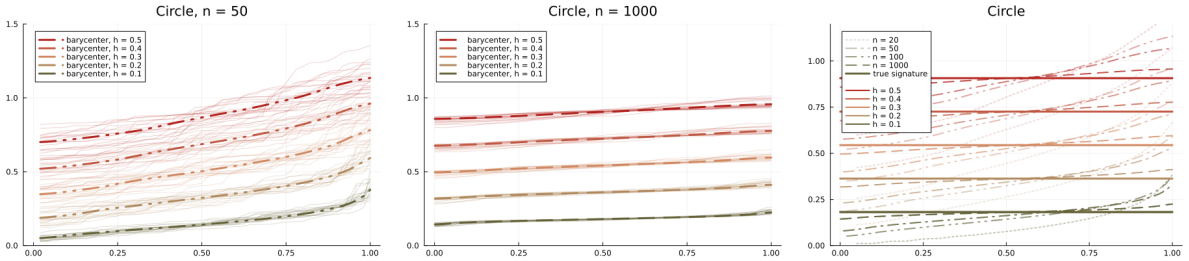


Figure 1: Signatures barycentres pour la loi uniforme sur le cercle pour des échantillons de taille 50 (à gauche), de taille 1000 (au milieu), de taille 20 à 1000 et la signature limite (*true signature*, à droite).

dans l'échantillon. Lorsque la taille de l'échantillon, n, tend vers l'infini, la signature converge vers $s_h(\mu_0) = \delta_{\mathrm{d}_h}$ où d_h est la valeur constante de la DTM de μ_0 qui dépend de h. Il est aussi possible de calculer le barycentre de la signature-DTM comme la loi de fonction quantile

$$F^- : x \in \mathbb{R} \mapsto \mathbb{E}_{\mu_0}\left[F^-_{s_h\left(\frac{1}{n}\sum_{j=1}^n \delta_{X_j}\right)}(x)\right], \tag{5}$$

donnée par l'espérance des fonctions quantiles des signatures aléatoires $s_h\left(\frac{1}{n}\sum_{j=1}^n \delta_{X_j}\right)$, pour des n-échantillons iid de loi μ_0. On note cette signature barycentre $\bar{s}_h(\hat{\mathbb{\mu}}_{0,n})$. Les signatures barycentre et signatures limite pour le cercle sont représentées dans la Figure 1.

2 Les tests statistiques

2.1 Introduction des tests statistiques

Étant donnée la mesure empirique $\boldsymbol{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ de loi $\mathbb{\mu}_n$ définie sur l'ensemble des mesures uniformes supportées sur n points $\mathscr{P}_n(\mathscr{X}) = \{\frac{1}{n}\sum_{i=1}^n \delta_{x_i},\, x_1, \ldots, x_n \in \mathscr{X}\}$, on définit deux familles de tests statistiques, $(\boldsymbol{\phi}^{\mathrm{hom}}_{n,h})_{0<h<1}$ et $(\boldsymbol{\phi}^{\mathrm{iid}}_{n,h})_{0<h<1}$, de l'hypothèse

$$\mathrm{H}_0 : \mathbb{\mu}_n = \hat{\mathbb{\mu}}_{0,n}$$

contre l'hypothèse alternative

$$\mathrm{H}_1 : \mathbb{\mu}_n \neq \hat{\mathbb{\mu}}_{0,n},$$

où $\hat{\mathbb{\mu}}_{0,n}$ est la loi de $\boldsymbol{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ lorsque $X_1, \ldots, X_n$ est un échantillon de n variables indépendantes de loi μ_0. On considère des alternatives de type:

$$\mathrm{H}_1(\mathrm{d}, \mu_0, \mathrm{c}) = \{\hat{\mathbb{\mu}}_{\mathrm{n}} \in \mathscr{P}(\mathscr{P}_{\mathrm{n}}(\mathscr{X})) \mid \mathrm{d}(\mu, \mu_0) > \mathrm{c}\}, \tag{6}$$

où la mesure $\mathbb{\mu}_n = \hat{\mathbb{\mu}}_n$ repose sur un échantillon de n variables indépendantes $(X_1, X_2, \ldots, X_n)$, de même loi $\mu \in \mathscr{P}(\mathscr{X})$.

La statistique de paramètre h de la famille de tests d'uniformité est donnée par :

$$\mathbf{T}^{\mathrm{hom}}_{n,h} = \mathcal{W}_1(s_h(\boldsymbol{\mu}_n), s_h(\mu_0)) \tag{7}$$

et le test

$$\boldsymbol{\phi}_{n,h}^{\text{hom}} = \mathbb{1}_{\mathbf{T}_{n,h}^{\text{hom}} > \text{q}_{1-\alpha,\text{n,h}}^{\text{hom}}}, \tag{8}$$

où $\text{q}_{1-\alpha,\text{n,h}}^{\text{hom}}$ est le quantile d'ordre $1-\alpha$ de $\mathbf{T}_{n,h}^{\text{hom}}$ lorsque $\mathbb{\mu}_n = \hat{\mathbb{\mu}}_{0,n}$, approché par Monte-Carlo.

La statistique de paramètre h de la famille de tests d'uniformité et d'indépendance est donnée par :

$$\mathbf{T}_{n,h}^{\text{iid}} = \mathcal{W}_2(s_h(\boldsymbol{\mu}_n), \bar{s}_h(\hat{\mathbb{\mu}}_{0,n})) \tag{9}$$

et le test

$$\boldsymbol{\phi}_{n,h}^{\text{iid}} = \mathbb{1}_{\mathbf{T}_{n,h}^{\text{iid}} > \text{q}_{1-\alpha,\text{n,h}}^{\text{iid}}}, \tag{10}$$

où $\text{q}_{1-\alpha,\text{n,h}}^{\text{iid}}$ est le quantile d'ordre $1-\alpha$ de $\mathbf{T}_{n,h}^{\text{iid}}$ lorsque $\mathbb{\mu}_n = \hat{\mathbb{\mu}}_{0,n}$, approché par Monte-Carlo.

2.2 Consistance des tests

Les tests $\boldsymbol{\phi}_{n,h}^{\text{hom}}$ et $\boldsymbol{\phi}_{n,h}^{\text{iid}}$ sont asymptotiquement uniformément consistants, contre des alternatives du type

$$\text{H}_1 \text{“}\mathcal{W}_1(s_h(\mu), s_h(\mu_0)) > \eta\text{”},$$

pour $\eta > 0$, au sens suivant.

Theorem 1 *Pour tout $\eta > 0$,*

$$\lim_{n\to\infty} \inf_{\mu\in\mathscr{P}(\mathscr{X}),\, \mathcal{W}_1(s_h(\mu),s_h(\mu_0))>\eta} \mathbb{P}_\mu(\mathbf{T}_{n,h}^{\text{hom}} > \text{q}_{1-\alpha,\text{n,h}}^{\text{hom}}) = 1 \tag{11}$$

et,

$$\lim_{n\to\infty} \inf_{\mu\in\mathscr{P}(\mathscr{X}),\, \mathcal{W}_2(s_h(\mu),s_h(\mu_0))>\eta} \mathbb{P}_\mu(\mathbf{T}_{n,h}^{\text{iid}} > \text{q}_{1-\alpha,\text{n,h}}^{\text{iid}}) = 1. \tag{12}$$

3 Illustrations Numériques

3.1 Comparaison aux tests d'uniformité sur le cercle, la sphère, la Grassmanienne

Dans les Figures 2 et 3, nous comparons nos deux familles de tests aux méthodes classiques implémentées dans le package R de García-Portugués et al.(2024), pour le cercle et la sphère. En plus de faire varier le paramètre $h \in [0, 1]$ dans nos tests, nous considérons une version agrégée des tests dans laquelle le meilleur paramètre h est choisi de façon automatique à partir des données, la valeur critique est ensuite recalibrée pour que le test reste de niveau 5%. Nos méthodes avec un paramètre h proche de 1 détectent presque aussi bien que les meilleurs tests les alternatives de type loi von-Mises - Fisher avec un seul mode. Au contraire, nos méthodes avec un paramètre h proche de 0 détectent au moins aussi bien, voire mieux que les meilleurs tests, des alternatives de types mélanges de lois de type von-Mises - Fisher, avec des modes répartis de façon symétrique sur le cercle ou la sphère. Le même type de phénomène apparaît dans la Grassmanienne, dans la Figure 4, où la performance de nos tests $(\boldsymbol{\phi}_{n,h}^{\text{iid}})_h$ est comparée à la fonction `grassmann.utest` de la librairie R `Riemann`.

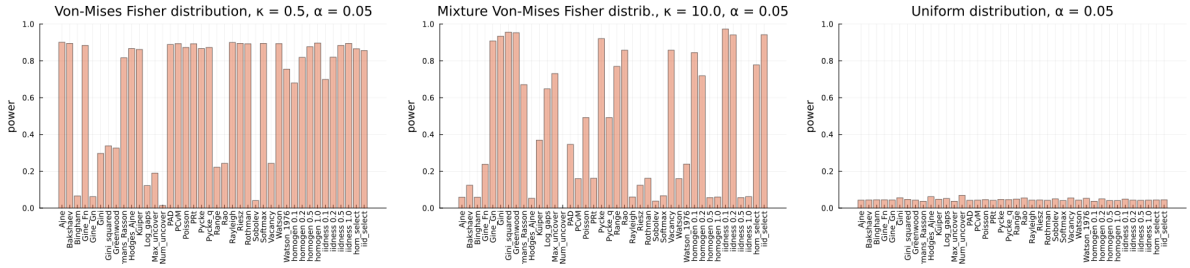


Figure 2: Puissance sur le cercle. Alternatives : unimodale, multimodale, loi uniforme

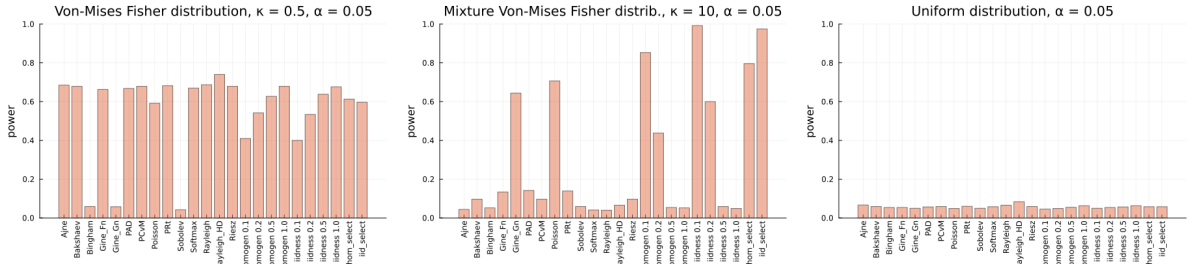


Figure 3: Puissance sur la sphère. Alternatives : unimodale, multimodale, loi uniforme

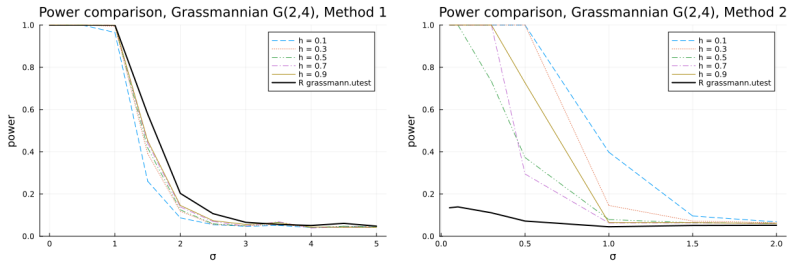


Figure 4: Puissance sur la Grassmannienne $\mathbb{G}(2,4)$. Méthode 1 : deux premiers vecteurs propres de la matrice de covariance de 50-échantillons $(s_i + X_i)_{1\leq i\leq 50}$ où $(s_i)_{1\leq i\leq 50} \in \mathbb{R}^4$ sont issues du jeu de données R `iris` et $(X_i)_{1\leq i\leq 50}$ sont des 50-échantillons iid de loi normale $\mathcal{N}(0,\sigma)$. Méthode 2 : mélange de 6 lois normales dans $\mathbb{G}(2,4)$ (définie à partir de la fonction exponentielle et logarithme), de centres (e_1,e_2), (e_1,e_3), (e_1,e_4), (e_2,e_3), (e_2,e_4), (e_3,e_4).

3.2 Comparaison des puissances pour des échantillons non iid

Les deux familles de tests agissent de façon contraire lorsque les données $(X_1,\ldots,X_n)$ n'ont pas été générées de façon indépendante. On considère des points issus d'une grille régulière sur la sphère, auxquels on a ajouté un bruit de type von Mises, avec un paramètre κ croissant. (Pour $\kappa = 0$, on a un échantillon iid de loi μ_0 sur la sphère. Pour $\kappa \to +\infty$, on a une grille régulière) D'après la Figure 5, les tests d'uniformité $\boldsymbol{\phi}^{\text{hom}}_{n,h}$ ne rejettent jamais H_0 pour la grille. En effet, les points sont au moins aussi uniformément répartis sur la sphère que le serait un échantillon iid de loi uniforme. Au contraire, les tests d'uniformité et d'indépendance $\boldsymbol{\phi}^{\text{iid}}_{n,h}$ rejettent systématiquement H_0, car les points ne sont pas indépendants.

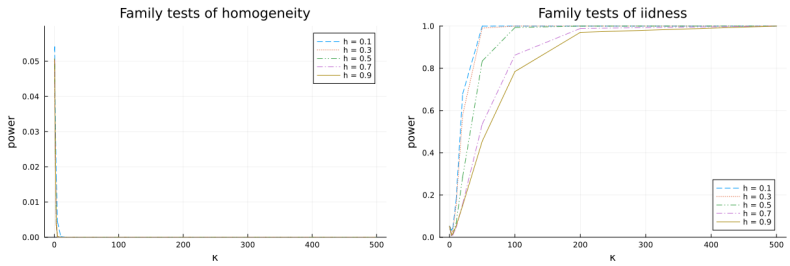


Figure 5: Comparaison des puissances des deux familles de tests

4 Conclusion et Perspectives

Pour de grands paramètres h, les tests se comparent aux meilleurs tests de la littérature, pour rejeter des alternatives unimodales, alors que pour de petits paramètres h, les tests montrent de meilleures performances que les tests existants sur des alternatives multimodales symétriques. Les tests d'uniformité et indépendance pourront par exemple être utilisés pour tester des méthodes déterministes de génération d'échantillons indépendants de loi uniforme.

Bibliographie

Brécheteau, C. (2019), A statistical test of isomorphism between metric-measure spaces using the distance-to-a-measure signature, *Electronic Journal of Statistics* Vol. 13, No. 1, 795-849

Christensen, J.P.R. (1970), On some measures analogous to Haar measure, *Mathematica Scandinavica* Vol. 26, 103-106

Loomis, L.H. (1945), Abstract congruence and the uniqueness of Haar measure, *Annals of Mathematics. Second Series* Vol. 46, 348-355

Chazal, F. and Cohen-Steiner, D. and Mérigot, Q. (2010), Geometric inference for probability measures, *Found. Comput. Math.* Vol. 11, 733-751

Hallin, M. and Mordant, G. and Segers, J. (2021), Multivariate goodness-of-fit tests based on Wasserstein distance, *Electron. J. Stat.* Vol. 15, 1328–1371

García-Portugués, E. and Verdebout, T. (2018), An overview of uniformity tests on the hypersphere.

García-Portugués, E. and Verdebout, T. (2024), sphunif: Uniformity Tests on the Circle, Sphere, and Hypersphere, *R package version 1.4.0*, `https://CRAN.R-project.org/package=sphunif`.

Chikuse, Y. (2003), Statistics on special manifolds, *Lecture Notes in Statistics, Springer-Verlag, New York*

Estimation non-paramétrique pour des données circulaires censurées

Nicolas Conanec [1] & Claire Lacour[2] & Thanh Mai Pham Ngoc[1]

[1] *LAGA (UMR 7539), Université Sorbonne Paris Nord, Institut Galilée, 99 avenue Jean-Baptiste Clement, 93430 Villetaneuse, France, {conanec, phamngoc}@math.univ-paris13.fr*

[2] *Université Gustave Eiffel, Univ Paris Est Creteil, CNRS, LAMA UMR8050 F-77447 Marne-la-Vallée, France, claire.lacour@univ-eiffel.fr*

Résumé. Nous nous intéressons à un problème d'estimation de densité de probabilité pour des données circulaire censurées. Notre modèle de censure est le suivant, X est la variable aléatoire d'intêret à valeurs dans le cercle $\mathbb{S}^1$ et admettant pour densité f, la fonction que l'on souhaite estimer, (L, U) est un couple de variable aléatoires à valurs dans $(\mathbb{S}^1)^2$ tel que X est indépendant du couple (L, U) et on observe effectivement X si $X \in [L, U]$ et on ne l'observe pas sinon (on choisit alors une valeur arbitraire hors de $[0, 2\pi[$ pour ces cas-là, ici on prend $-\pi$). Si on note $\Delta = \mathbb{1}_{\{X\in[L,U]\}}$ et $X' = \Delta X - \pi(1-\Delta)$ alors nos observations sont un n-échantillon de triplets (X'_i, L_i, U_i). Nous estimons les deux fonctions suivantes $\phi : x \in \mathbb{S}^1 \mapsto \mathbb{P}(x \in [L, U])$ et $\psi = f\phi$. On estime ϕ par son estimateur empirique et ψ, que l'on supposera appartenant à une classe de Sobolev, par minimum de contraste sur le sous espace S_m de $\mathbb{L}^2(\mathbb{S}^1)$ des polynômes trigonométriques de degré m. De ces deux estimateurs on définit un estimateur quotient de f. Nous montrons que la vitesse de convergence du risque quadratique de cet estimateur est la vitesse non-paramétrique classique. Ensuite nous démontrons un résultat de borne inférieure pour le risque intégré qui assure l'optimalité de notre estimateur quand ψ est au moins aussi régulière que f. De plus, nous pouvons mettre en place une procédure adaptative pour l'estimation de ψ, qui estime la dimension de l'espace S_m, et ce nouvel estimateur conserve sa vitesse de convergence.

Mots-clés. Estimation non-paramétrique, Données circulaires, Données censurées, Estimation adaptative, Borne inférieure

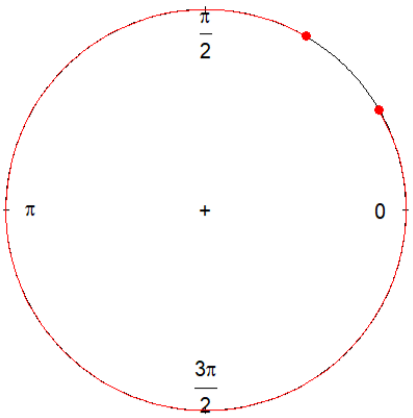
Abstract. We are interested in a density estimation problem for censored circular data. Our censorship model is the following, X is a random variable with values in the circle $\mathbb{S}^1$ with probability density function f to estimate, (L, U) is a couple of random variables with values in $(\mathbb{S}^1)^2$ such that X is independent from the couple (L, U) and X is observed if $X \in [L, U]$ and we do not observe it otherwise (we choose an arbitrary value outside of $[0, 2\pi[$ in those case, here we took $-\pi$). We write $\Delta = \mathbb{1}_{\{X\in[L,U]\}}$ and $X' = \Delta X - \pi(1-\Delta)$. Then our observations are a sample of n iid triplets (X'_i, L_i, U_i). We estimate the following functions $\phi : x \in \mathbb{S}^1 \mapsto \mathbb{P}(x \in [L, U])$ and $\psi = f\phi$. We estimate ϕ with its empirical counterpart and ψ, that we will assume it belongs to a Sobolev class, with minimum contrast on the subspace S_m of $\mathbb{L}^2(\mathbb{S}^1)$ of the trigonometric polynomials of degree m. With those estimators we construct a quotient estimator of f. We show that the rate of convergence of the mean squared error of the estimator is the classic nonparametric rate. Moreover we proved a lower

bound result for the the integrated risk which ensures the optimality of our estimator when ψ is as regular as f. Finally we showed that we can use an adaptive method in the estimation of ψ, which estimates the dimension of the space S_m, and this new estimator has the same rate of convergence as before.

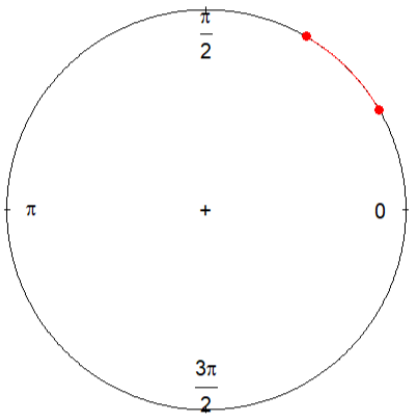
Keywords. Nonparametric estimation, Circular data, Censored data, Adaptive estimator, Lower bound

1 Données circulaires et modèle de censure

On parle de données directionnelles quand les observations sont des éléments de $\mathbb{R}^d$ pour $d \geq 2$ et quand seule la direction importe, on peut donc supposer le vecteur de norme 1. Ces données peuvent être observées dans de nombreux contextes, en météorologie si on s'intéresse à la direction du vent ou les instant d'apparition de tornades; en biologie si on s'intéresse aux directions de migration animale; en sciences de la terre si on veut étudier les épicentres de séismes ou encore en médecine pour la vectocardiographie. Cependant les données directionnelles ne se manipulent pas exactement comme des données réelles. Ici nous nous focaliserons sur des données circulaires, i.e des élements de $\mathbb{S}^1$. Nous voulons réaliser une estimation non-paramétrique de la densité de probabilité d'une variable aléatoire circulaire dont l'observation est censurée.
Avant de définir notre modèle de censure nous devons discuter de l'aspect circulaire de nos données, en particulier la définition d'intervalle circulaire. Après avoir choisi une direction initiale (le 0), et une orientation (le sens trigonométrique ici), nous pouvons représenter le cercle $\mathbb{S}^1$ par le segment $[0, 2\pi[$ de $\mathbb{R}$. Si $x, y \in \mathbb{S}^1$, le segment de x vers y suivant une orientation et celui de x vers y suivant l'autre orientation ne sont pas les mêmes, ils sont complémentaires au sens où ils sont disjoints et leur réunion forme le cercle entier. Ainsi on va noter $[x, y]$ le segment sur le cercle des points entre x et y suivant l'orientation que l'on a choisie.



(a) Intervalle circulaire $\left[\frac{\pi}{3}, \frac{\pi}{6}\right]$.



(b) Intervalle circulaire $\left[\frac{\pi}{6}, \frac{\pi}{3}\right]$.

Le modèle de censure que l'on considère se rapproche de celui étudié par Turnbull (1974) à la différence qu'ici dans le cas d'une donnée censurée nous observons les deux éléments de censure. Le modèle de censure que nous allons étudier ici est un modèle introduit par Jammalamadaka et Mangalam (2009). Soit X une variable aléatoire circulaire de densité de probabilité f et (L, U) un couple de variable de censure tel que X est indépendant de (L, U). Le segment $[L, U]$ est appelé fenêtre d'observation car nous observons X si $X \in [L, U]$. Au contraire $]U, L[$ est l'intervalle de censure car nous n'observons pas X si $X \in]U, L[$, dans ce cas nous observons une valeur arbitraire hors de $[0, 2\pi[$, nous avons choisi $-\pi$. Ainsi on définit

$$\begin{cases} X' = \begin{cases} X \text{ , si } X \in [L, U] \\ -\pi \text{ , sinon.} \end{cases} \\ \Delta = \mathbb{1}_{\{X \in [L,U]\}} \end{cases}$$

On peut également écrire $X' = X\Delta - \pi(1-\Delta)$. On observe un n-échantillon $(X'_i, L_i, U_i)_{i \in [\![1,n]\!]}$. Notre but est d'estimer la densité de probabilité f à partir de ces observations.

2 Estimateur

On s'inspire de la méthode d'estimation utilisée par Brunel et Comte (2009). De par la nature circulaire de nos données nous allons travailler sur $\mathbb{L}^2(\mathbb{S}^1)$, espace que l'on va munir du produit scalaire $< f, g >= \int_{\mathbb{S}^1} f(x)g(x)\mathrm{d}x$ et de la norme associée notée $\|\cdot\|_2$. On remarque que si t est une fonction quelconque de $\mathbb{L}^2(\mathbb{S}^1)$

$$\begin{aligned} \mathbb{E}(\Delta t(X')) &= \int_{\mathbb{S}^1}\int_{\mathbb{S}^1}\int_{\mathbb{S}^1} \mathbb{1}_{\{x \in [l,u]\}} t(x) f(x) \mathbb{P}_{(L,U)}(\mathrm{d}l, \mathrm{d}u)\mathrm{d}x \\ &= \int_{\mathbb{S}^1} t(x)f(x) \underbrace{\int_{\mathbb{S}^1}\int_{\mathbb{S}^1} \mathbb{1}_{\{x \in [l,u]\}} \mathbb{P}_{(L,U)}(\mathrm{d}l, \mathrm{d}u)}_{:=\phi(x)} \mathrm{d}x \\ &= \int_{\mathbb{S}^1} t(x)f(x)\phi(x)\mathrm{d}x \\ &=< t, f\phi >=< t, \psi >, \end{aligned} \tag{1}$$

si on pose $\psi = f\phi$. Ainsi en estimant ψ et ϕ le quotient des deux estimateurs estimera f.

Pour l'estimation de ψ l'égalité (1) justifie l'usage d'un estimateur par minimum de contraste. Puisque nous travaillons sur le cercle les fonctions trigonométrique semblent naturelles pour réaliser l'estimation de nos fonctions. Ainsi on note S_m le sous espace de $\mathbb{L}^2(\mathbb{S}^1)$ des polynômes trigonométriques de degré m, i.e. généré par les fonctions $\{\varphi_0 = \frac{1}{\sqrt{2\pi}}, \varphi_{2j-1} = \frac{1}{\sqrt{\pi}}\cos(jx), \varphi_{2j} = \frac{1}{\sqrt{\pi}}\sin(jx)| \text{ pour } j \in [\![1, m]\!]\}$. Cet espace de dimension $D_m = 2m + 1$ est l'espace dans lequel on définit notre estimateur, ainsi le choix de m qui définit ce sous espace est une question importante. On définit le contraste suivant

$$\gamma_n^{\psi} : t \in \mathbb{L}^2(\mathbb{S}^1) \mapsto \|t\|_2^2 - \frac{2}{n}\sum_{i=1}^{n} \Delta_i t(X_i).$$

Ainsi un estimateur de ψ est

$$\hat{\psi}_m = \underset{t \in S_m}{\operatorname{argmin}}\, \gamma_n^{\psi}(t). \tag{2}$$

La justification de ce contraste vient de l'égalité (1), on a pour tout $t \in \mathbb{L}^2([0, 2\pi])$,

$$\mathbb{E}(\gamma_n^{\psi}(t)) = \|t\|_2^2 - 2 < t, \psi >= \|t - \psi\|_2^2 - \|\psi\|_2^2.$$

Comme $\hat{\psi}_m$ minimise le contraste γ_n^{ψ}, par l'égalité précédente il est également suceptible de minimiser la norme $\|t - \psi\|_2^2$ sur S_m et donc être un estimateur pertinent de ψ. En ce qui concerne l'estimation de ϕ, on remarque que l'on peut également la noter

$$\phi : x \in \mathbb{S}^1 \mapsto \mathbb{P}(x \in [L, U]).$$

Comme la fonction est définie comme une probabilité, on peut utiliser son estimateur empirique

$$\hat{\phi} : x \in \mathbb{S}^1 \mapsto \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x \in [L_i, U_i]\}}.$$

Enfin après le passage au quotient, puisque l'on veut estimer une densité, on applique une partie positive et un seuil pour limiter les effets de bord si $\hat{\phi}$ est trop proche de 0. Ainsi notre estimateur de f est défini comme

$$\hat{f} : x \mapsto \left(\frac{\hat{\psi}_m(x)}{\hat{\phi}(x) \vee n^{-1/2}} \right)_+. \tag{3}$$

3 Résultats

On suppose qu'il existe $\phi_0 > 0$ tel que pour tout $x \in \mathbb{S}^1, \phi(x) \geq \phi_0$.

Une étude du MISE de $\hat{f}$ nous donne l'inégalité suivante

Théorème 1. *Avec le modèle décrit en Section 1 et l'estimateur $\hat{f}$ défini en (3), on a la majoration suivante du risque quadratique intégré,*

$$\mathbb{E}(\|\hat{f} - f\|_2^2) \leq \frac{1}{\phi_0^2} \left(4\mathbb{E}(\|\hat{\psi}_m - \psi\|_2^2) + \frac{\|f\|_2^2}{n} \right) + \left(\frac{1}{2\pi} + \|\psi\|_2^2 + 4\|f\|_2^2 \right) 6ne^{-2n\frac{\phi_0^2}{36}},$$

et en ce qui concerne le MISE de $\hat{\psi}_m$ on obtient

Théorème 2. *Avec le modèle décrit en Section 1, ψ_m la projection de ψ sur S_m et l'estimateur $\hat{\psi}_m$ défini en (2), on a la majoration suivante du risque quadratique intégré,*

$$\mathbb{E}(\|\psi - \hat{\psi}_m\|_2^2) \leq \|\psi - \psi_m\|_2^2 + \frac{D_m \mathbb{E}(\Delta_1)}{2\pi n}.$$

En supposant de plus que ψ appartient à une classe de Sobolev $W(\beta, L)$ où β est la régularité des fonctions, et en choisissant une dimension D_m de l'ordre de $n^{\frac{1}{2\beta+1}}$ on obtient que la vitesse de convergence du MISE de $\hat{\psi}_m$ est alors la vitesse non-paramétrique classique et on déduit alors celle de $\hat{f}$.

Corollaire 1. *Soit S_m l'espace des polynômes trigonométriques de degré m et supposons que ψ est un élément de $W(\beta, L)$ sur $\mathbb{S}^1$ avec $\beta > 0$. De plus si m est choisi tel que $D_m = cn^{\frac{1}{2\beta+1}}$, avec $c > 0$ constante, alors on a la majoration suivante,*

$$\mathbb{E}(\|\hat{f} - f\|_2^2) = \mathcal{O}\left(n^{\frac{-2\beta}{2\beta+1}}\right).$$

Cependant ce dernier résultat suppose que l'on sache choisir la bonne valeur de m mais cela n'est pas possible en pratique car β est inconnu. Pour régler ce problème on utilise une méthode de sélection de modèles par minimum de contraste pénalisé.

Théorème 3. *On pose*

$$\hat{m} = \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \left(\gamma_n^{\psi}(\hat{\psi}_m) + pen(m)\right),$$

où $pen(m) = \frac{\kappa}{2\pi}\frac{D_m}{n}\left(\frac{1}{n}\sum_{i=1}^n \Delta_i\right)$, et κ une constante universelle ($\kappa > 8$ convient), on obtient alors l'inégalité suivante

$$\mathbb{E}\left(\|\hat{\psi}_{\hat{m}} - \psi\|_2^2\right) \leq C \inf_{m \in \mathcal{M}_n}\left(\|\psi - \psi_m\|_2^2 + \frac{D_m \mathbb{E}(\Delta_1)}{2\pi n}\right) + \frac{C'}{n},$$

pour une constante C qui ne dépend que de κ et C' une constante qui dépend uniquement du modèle.

En d'autres termes notre estimateur de ψ est adaptatif, de plus l'ajout de l'estimation de m ne dégrade pas la vitesse de convergence ni de ψ ni de f.

Enfin nous avons étudié la borne inférieure du risque de ce problème d'estimation. Si on suppose également que f est une fonction de Sobolev de régularité $\beta > 1/2$ alors on a démontré le résultat suivant.

Théorème 4. *Supposons $\beta > \frac{1}{2}$ et $L > 0$. Avec le modèle décrit en Section 1 on a la minoration suivante,*

$$\liminf_{n \to +\infty} \inf_{\tilde{f}} \sup_{f \in W(\beta, L)} \mathbb{E}_f\left(n^{\frac{2\beta}{2\beta+1}} \|\tilde{f} - f\|_2^2\right) \geq c,$$

où $\inf_{\tilde{f}}$ est l'infimum sur tous les estimateurs possibles et la constante c ne dépend que des paramètres β et L.

4 Simulations

Dans Jammalamadaka et Mangalam (2009) les auteurs ont réalisé deux estimations paramétriques, la moyenne $\mu = 2$ et le paramètre de concentration $k = 1$ d'une loi de Von Mises $M(\mu, k)$. Nous avons répliqué leur méthode de simulation et réalisé l'estimation non-paramétrique de la densité, de cela nous avons déduit des estimations de ces paramètres et obtenu des résultats semblables et ce malgré l'absence totale d'information sur la loi.

Nous réalisons des simulations de notre estimateur, en utilisant notamment le package CAPUSCHE pour la selection de modèles. Comme le montre la Figure 3, même avec 46% de données censurées l'estimation que l'on obtient est satisfaisante.

Longueur de l'arc de censure	Estimation de μ			Estimation de k		
	$\tilde{\mu}$	$\breve{\mu}$	$\hat{\mu}$	$\tilde{k}$	$\breve{k}$	$\hat{k}$
1	1.985	1.985	2.005	1.047	1.046	1.041
3	2.006	2.003	1.990	1.037	1.031	1.021

Figure 2: Simulations, avec un échantillon de 200 valeurs simulées, de $\mu = 2$ et $k = 1$ quand X suit une $M(\mu, k)$, L suit la loi uniforme circulaire et $U = L + l$ où l est la longueur de l'arc de censure. Nos estimations sont $\hat{\mu}$ et $\hat{k}$, les autres sont celles de Jammalamadaka et Mangalam.

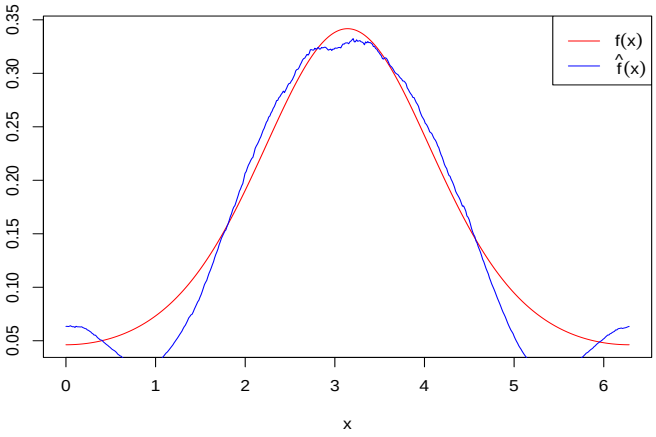


Figure 3: Simulation de l'estimation de la densité de la $M(\pi, 1)$ quand L suit la $M\left(\frac{2\pi}{3}, 1\right)$ et U suit la $M\left(\frac{4\pi}{3}, 1\right)$ avec un échantillon de $n = 1000$ données. L'estimation est en bleu et la vraie densité f en rouge. Sur l'ensemble de l'échantillon, 45.9% des données étaient censurées.

Bibliographie

Brunel E. et Comte F. (2009), Cumulative Distribution Function Estimation under Interval Censoring Case 1, *Electronic Journal of Statistics*, 3, pp. 1-24.

Jammalamadaka S. Rao et Mangalam V. (2009), A General Censoring Scheme for Circular Data , *Statistical Methodology*, 6, pp. 280-289

Turnbull B. (1974), Nonparametric Estimation of a Survivorship Function with Doubly Censored Data, *Journal of the American Statistical Association*, 69, pp. 169-173

Test d'adéquation par moments trigonométriques pour K échantillons circulaires

Maxime Boucher [1] & Simos Meintanis [2]& Vivien Meurice & Thomas Verdebout [3]

[1] *Université de Namur, Namur, Belgique, maxime.boucher@unamur.be*
[2] *National and Kapodistrian University of Athens, Athens, Greece, simosmei@econ.uoa.gr*
[3] *Université libre de Bruxelles, Bruxelles, Belgique, vivien.meurice@ulb.be*
[4] *Université libre de Bruxelles, Bruxelles, Belgique, thomas.verdebout@ulb.be*

Résumé. Dans cette présentation, on propose un test d'adéquation, adapté à un cadre de données circulaires, pour K échantillons indépendants. On montre que ce test se construit sur la base des moments trigonométriques et qu'il peut être réécrit comme un test de Wald (test de contraintes linéaires). De plus, on peut évaluer les puissances (explicites) de ce test dans des cas où les données sont Von-Mises avec des perturbations locales pour en déduire les lois asymptotiques (chi-deux) de la statistique de test générale. A la suite de simulations Monte-Carlo pour illustrer nos résultats et les puissances explicites de notre test, on propose également une première application sur l'étude des longitudes des taches solaires au cours des différents cycles solaires.

Mots-clés. Données circulaires, test d'adéquation, moments trigonométriques, test d'hypothèse, statistiques most-stringent.

Abstract. In this presentation, we propose a goodness-of-fit test adapted to a framework of circular data for K independent samples. We show that this test is constructed based on trigonometric moments and can be rewritten as a Wald test (linear constraint test). Furthermore, we can assess the power of this test in cases where the data follow a Von Mises distribution with local perturbations to derive the asymptotic distributions (chi-squared) of the general test statistics. Following Monte Carlo simulations to illustrate our results and the explicit power of our test, we also present a first application to the study of the longitudes of solar spots over different solar cycles.

Keywords. Circular data, goodness-of-fit, trigonometric moments, hypothesis testing, most-stringent test statistics.

Test d'adéquation pour K échantillons circulaires

Introduction : données directionnelles et circulaires

Depuis les dernières décennies, l'étude des statistiques directionnelles a connu un essor important. On parle de données directionnelles lorsque l'on dispose d'observations sur la sphère de dimension p, noté $\mathcal{S}^{p-1} = \{x \in \mathbb{R}^p | x'x = 1\}$ où on désigne ici la transposée du vecteur x par x'. Dans le cas particulier où $p = 2$, on parle de données circulaires. Les données sont alors disposées sur le cercle unité $\mathcal{S}^1$. Dans le cas de données circulaires (et directionnelles), nous devons prendre en compte la géométrie du contexte avec par exemple les notions de moyennes empiriques (puisque si $X_1, \dots, X_n \in \mathcal{S}^{p-1}$, alors $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \notin \mathcal{S}^{p-1}$). Pour plus de détails sur les données directionnelles, on renvoie le lecteur vers [5].

Dans ce travail, nous proposons l'étude d'un test d'adéquation dans le cas particulier de données circulaires.

Problématique : test d'adéquation pour K variables circulaires

On suppose que l'on dispose de K variables aléatoires circulaires, de distribution sur le cercle unité $\mathcal{S}^1$, que l'on note $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(K)}$. On s'intéresse à savoir si la loi de probabilité pour toutes les variables aléatoires est la même. On notera $\theta^{(0)}$ la variable aléatoire qui a pour distribution la loi commune de toutes les autres. Autrement dit, on s'intéresse au test :

$$H_0 : "\theta^{(1)} \overset{\mathcal{L}}{=} \dots \overset{\mathcal{L}}{=} \theta^{(K)} \overset{\mathcal{L}}{=} \theta^{(0)}" \qquad \text{versus} \qquad H_1 : "\exists i \in [\![1, K]\!], \theta^{(i)} \overset{\mathcal{L}}{\neq} \theta^{(0)}$$

Pour tester l'adéquation entre les variables, on peut utiliser les fonctions de répartition. Dans ce travail, nous proposons d'utiliser les fonctions caractéristiques. Pour cela, on rappelle la définition de la fonction caractéristique de $\theta^{(k)}$:

$$\forall x \in \mathbb{R}, \phi^{(k)}(x) := \mathbb{E}\left[e^{ix\theta^{(k)}}\right] = \mathbb{E}\left[\cos\left(x\theta^{(k)}\right)\right] + i\mathbb{E}\left[\sin\left(x\theta^{(k)}\right)\right].$$

On peut alors réécrire le problème de test d'adéquation comme étant :

$H_0 : "\forall x \in \mathbb{R}, \phi^{(1)}(x) = \dots = \phi^{(K)}(x) = \phi^{(0)}(x)"$ versus $H_1 : "\exists i \in [\![1, K]\!], \phi^{(i)}(x) \neq \phi^{(0)}(x)"$.

Dans le cadre de données circulaires, par des arguments de symétrie, la fonction caractéristique ne doit être étudiée que pour des valeurs entières (voir [3] pour plus d'informations). Le problème revient donc à considérer le test :

$H_0 : "\forall m \in \mathbb{Z}, \phi^{(1)}(m) = \dots = \phi^{(K)}(m) = \phi^{(0)}(m)"$ vs $H_1 : "\exists i \in [\![1, K]\!], \phi^{(i)}(m) \neq \phi^{(0)}(m)"$.

Finalement, en considérant les moments trigonométriques des variables aléatoires $\theta^{(k)}$ pour $k = 1, \ldots, K$, et pour un entier $M \in \mathbb{N}^*$ fixé, regroupés dans les vecteurs

$$\mu_M^{(k)} := \left(\mathbb{E}\left[\cos\left(\theta^{(k)}\right)\right], \ldots, \mathbb{E}\left[\cos\left(M\theta^{(k)}\right)\right], \mathbb{E}\left[\sin\left(\theta^{(k)}\right)\right], \ldots, \mathbb{E}\left[\sin\left(M\theta^{(k)}\right)\right]\right)' \in \mathbb{R}^{2M}.$$

Le test peut s'écrire comme une égalité des moments trigonométriques de toutes les populations contre une alternative où au moins un vecteur $\mu_M^{(k)}$ diffère des autres. En réunissant tous les vecteurs en un avec la notation

$$\mu_M := \left(\mu_M^{(1)}, \ldots, \mu_M^{(K)}\right)' \in \mathbb{R}^{2MK},$$

le test se réécrit comme étant :

$$H_0 : "\mu \in \mathcal{M}(\gamma)" \qquad \text{vs} \qquad H_1 : "\mu \notin \mathcal{M}(\gamma)",$$

où la matrice $\gamma := (\mathbf{I_{2M}}, \ldots, \mathbf{I_{2M}})' \in \mathbb{R}^{2M \times 2MK}$ et où $\mathcal{M}(\gamma)$ est l'espace engendré par les colonnes de la matrice γ. Le problème d'adéquation devient alors un problème de contraintes linéaires : un test de Wald.

Résultats principaux

On dispose d'observations, sous la nulle,

$$\theta_1^{(1)}, \ldots, \theta_{n_1}^{(1)}, \ldots, \theta_1^{(K)}, \ldots, \theta_{n_K}^{(K)} \overset{i.i.d}{\sim} \theta^{(0)}$$

où on suppose que $n = \sum_{i=1}^{K} n_i$ et que $r_{i,n} = \frac{n_i}{n} \underset{n \to +\infty}{\longrightarrow} r_i \in]0, 1[$. On définit alors la version empirique de μ_M, notée $\hat{\mu}_M$, par

$$\hat{\mu}_M = \left(\hat{\mu}_M^{(1)}, \ldots, \hat{\mu}_M^{(K)}\right)',$$

où $\hat{\mu}_M^{(k)} := \frac{1}{n_k} \sum_{i=1}^{n_k} U_i^{(k)}$ avec $U_i^{(k)} := \left(\cos(\theta_i^{(k)}), \ldots, \cos(M\theta_i^{(k)}), \sin(\theta_i^{(k)}), \ldots, \sin(M\theta_i^{(k)})\right)'$. Par la Loi des Grands Nombres (LGN) et Théorème Central Limite (TCL), on sait que $\hat{\mu}_M$ est un estimateur consistant et asymptotiquement normal. On construit la matrice $P_\gamma := \mathbf{I_{2MK}} - \frac{\mathbf{1}}{\mathbf{K}} \gamma\gamma'$ de projection sur l'orthogonale de $\mathcal{M}(\gamma)$. De plus, on note $\Sigma^{(0)} = cov\left(U_1^{(1)}\right) \in \mathbb{R}^{2M \times 2M}$ la matrice de covariance, sous l'hypothèse nulle, du vecteur $U_1^{(1)}$. On définit ainsi :

$$D^{(0)} := \mathbf{I}_K \otimes \Sigma^{(0)} \in \mathbb{R}^{2MK \times 2MK}$$

Théorème 1. *Sous H_0, le test Φ_{st} rejette l'hypothèse nulle quand*

$$Q_{st} := \left\| P_\gamma \left(D^{(0)}\right)^{-1/2} \hat{\mu}_M \right\|^2 > \chi^2_{2M(K-1),1-\alpha},$$

où $\chi^2_{2M(K-1),1-\alpha}$ *est le quantile de loi du chi-deux à* $2M(K-1)$ *degrés de liberté à l'ordre* $1-\alpha$.

Remarque : Le test ici est non paramétrique. On ne fait pas d'hypothèse sur les distributions circulaires. On montre également que l'on peut transformer la statistique de test pour la rendre invariante par rotation au prix d'une estimation du paramètre de localisation éventuel θ_0. De plus, dans le théorème 1, on montre que le test reste valide si l'on estime la matrice de covariance $\Sigma^{(0)}$. Ces deux estimations doivent vérifier des propriétés de $\sqrt{n}$-consistence pour que le théorème (version estimée) reste valide.

Dans ce travail, on propose également d'étudier la puissance de notre test dans un cas particulier. On s'intéresse au cas où, sous l'hypothèse nulle :

$$\theta^{(0)} \sim VM\left(\lambda_0, \omega_0\right)$$

où $VM(\lambda_0, \omega_0)$ est la distribution de Von-Mises avec un paramètre de concentration $\lambda_0 > 0$ et un paramètre de localisation $\omega_0 \in \mathcal{S}^1$ (la loi de Von-Mises est l'équivalent gaussien dans le contexte directionnel). Sous l'hypothèse alternative H_1, on va considérer une perturbation de l'hypothèse nulle où, pour $k = 1, \ldots, K$,

$$\theta^{(k)} \sim VM\left(\lambda_0 + \tfrac{1}{\sqrt{n}}\ell_n^{(k)}, \omega_0 + \tfrac{1}{\sqrt{n}}\tau_n^{(k)}\right)$$

où $\left(\ell_n^{(k)}\right)_{n\geqslant 1}$ et $\left(\tau_n^{(k)}\right)_{n\geqslant 1}$ sont des suites convergentes vers des constantes non nulles $\ell^{(k)}$ et $\tau^{(k)}$ pour $k = 1, \ldots, K$. Ici, on a donc le test

$$H_0 : \text{"}\theta^{(k)} \overset{\mathcal{L}}{=} \theta^{(0)} \sim VM\left(\lambda_0, \omega_0\right) \text{ pour tout } k\text{"}$$

vs

$$H_1 : \text{"}\theta^{(k)} \sim VM\left(\lambda_0 + \tfrac{1}{\sqrt{n}}\ell_n^{(k)}, \omega_0 + \tfrac{1}{\sqrt{n}}\tau_n^{(k)}\right) \text{ pour tout } k\text{"}.$$

On montre que l'étude du log-rapport des vraisemblances conduit à un résultat LAN (Local Asymptotic Normality) :

Théorème 2. *Sous* H_0*, pour chaque population,*

$$\Lambda_{\theta,\lambda}^{(k)} \;:=\; \log\left(\frac{d\mathbb{P}_1}{d\mathbb{P}_0}\right) = \left(\ell_n, \tau_n\right)\left(\Delta_{n,\lambda}^{(k)}, \Delta_{n,\theta}^{(k)}\right)' - \frac{1}{2}\left(\ell_n, \tau_n\right)\Gamma_{\lambda,\theta}\left(\ell_n, \tau_n\right)' + o^{\mathbb{P}}(1)$$

où

- $\Delta_{n,\lambda}^{(k)} := n_k^{-1/2}\sum_{i=1}^{n_k}\left(\cos\left(\theta_i^{(k)} - \omega_0\right) - \frac{I_0'(\lambda_0)}{I_0(\lambda_0)}\right)$ *: suite centrale pour le paramètre* λ_0.
- $\Delta_{n,\theta}^{(k)} := \lambda n_k^{-1/2}\sum_{i=1}^{n_k}\sin\left(\theta_i - \omega_0\right)$ *: suite centrale pour le paramètre* ω_0.

et

$$\left(\Delta_{n,\lambda}^{(k)}, \Delta_{n,\theta}^{(k)}\right)' \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_2\left(0, \Gamma_{\lambda,\theta}\right)$$

où $\Gamma_{\lambda,\theta}$ *est la matrice de covariance (explicite) du vecteur central.*

On en déduit le résultat suivant, pour toutes les populations :

Théorème 3. *Sous* H_0*,*

$$\log\left(\frac{dP_1}{dP_0}\right) = h_n'\Delta - \frac{1}{2}h_n'\Gamma_\Delta h_n + o^{\mathbb{P}}(1)$$

- $\Delta' = \left(\Delta_{n,\lambda}^{(1)}, \ldots, \Delta_{n,\lambda}^{(K)}, \Delta_{n,\theta}^{(1)}, \ldots, \Delta_{n,\theta}^{(K)}\right)$,
- $h_n' = \left(\ell_n^{(1)}, \ldots, \ell_n^{(K)}, \tau_n^{(1)}, \ldots, \tau_n^{(K)}\right)$,

avec

$$\Delta \xrightarrow[n\to+\infty]{} \mathcal{N}_{2K}\left(0, \Gamma_\Delta\right) \qquad \textit{et} \qquad h_n'\Delta \xrightarrow[n\to+\infty]{} \mathcal{N}_1\left(0, h'\Gamma_\Delta h\right)$$

On en déduit de ce résultat, à l'aide du troisième lemme de Le Cam, la puissance de notre test dans le cas particulier Von-Mises.

Théorème 4. *Sous* H_1*, la statistique* $\hat{Q}_{st} := \|\mathbf{P}_\gamma\left(\hat{D}^{(0)}\right)^{-1/2}\hat{\mu}_{\mathbf{M}}\|^2$ *suit asymptotiquement une loi du* χ^2 *avec* $2M(K-1)$ *degrés de liberté et un paramètre de décentralité (explicite)* $\nu = c'(D^{(0)})^{-1/2}\mathbf{P}_\gamma(D^{(0)})^{-1/2}c$*.*

Remarque : Le vecteur c correspond à la covariance, sous H_0 entre la statistique de test et la suite centrale. Ce vecteur est explicite. A l'aide du théorème 4, on dispose donc de la puissance du test sous l'alternative de perturbation de Von-Mises.

Illustration : Simulations Monte-Carlo

Nous appuyons nos travaux sur des simulations Monte-Carlo. Notamment, nous avons considéré le modèle suivant :

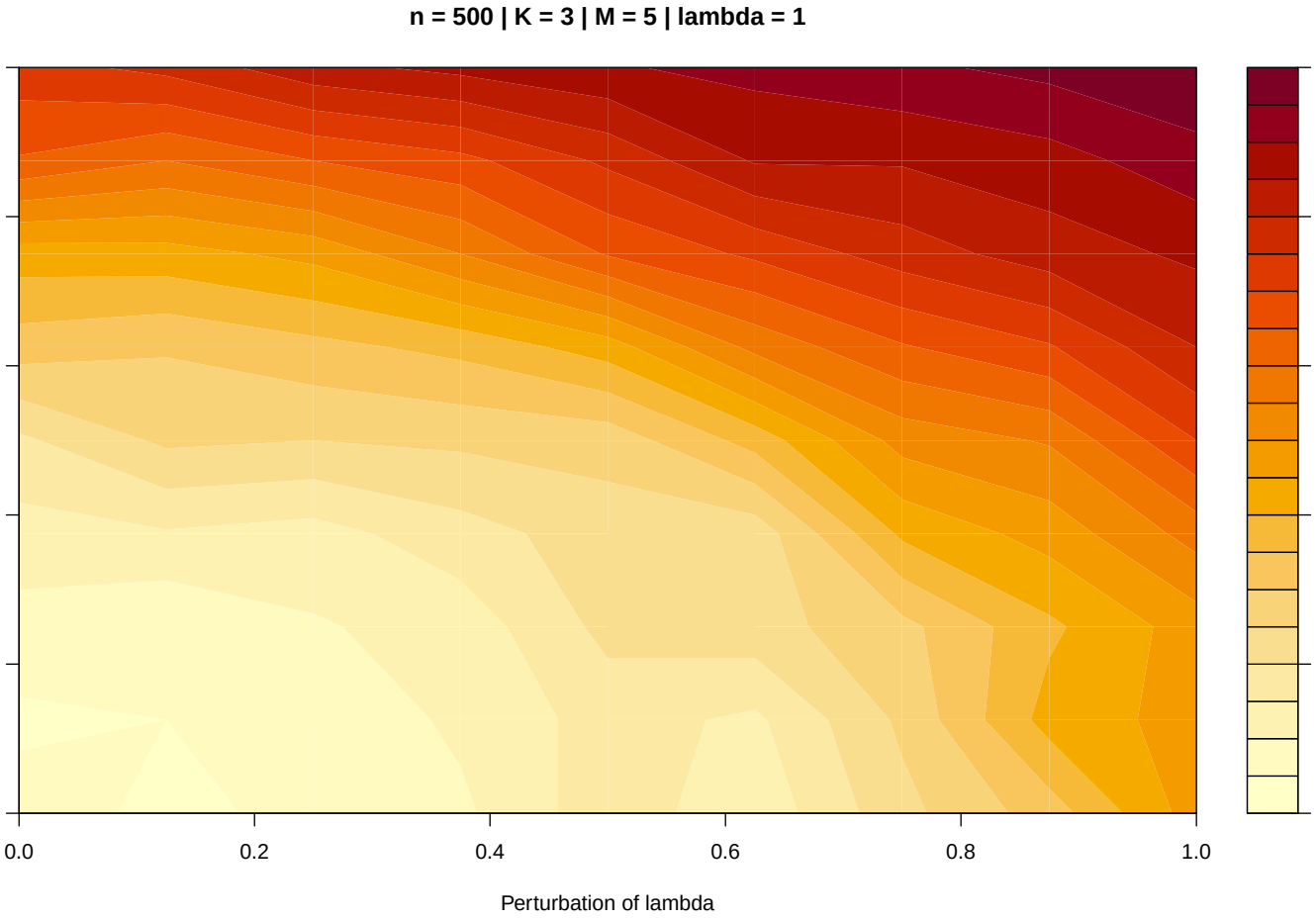
$$\theta_1^{(k)}, \ldots, \theta_{n_k}^{(k)} \overset{i.i.d}{\sim} VM\left(\lambda_0 + \tfrac{1}{\sqrt{n}}\ell_n^{(k)}, \omega_0 + \tfrac{1}{\sqrt{n}}\tau_n^{(k)}\right)$$

en choisissant, comme paramètres pour les simulations, avec $R = 500$ réplications :

- $K = 3$, $M = 5$, $n = 500$,
- $\lambda_0 = 1$, $\omega_0 = \frac{\pi}{4}$,

- $\tau_n^{(1)} = \ell_n^{(1)} = (0, 1, \ldots, 8)$,
- $\tau_n^{(k)} = \ell_n^{(k)} = 0$ pour $k = 2, \ldots, K$.

Dans cet exemple, sous l'alternative, seule la population 1 est perturbée. De plus, avec ce choix pour les suites de perturbations sur les paramètres de localisation et de concentration, l'hypothèse nulle correspond au cas où $\ell_n^{(1)} = \tau_n^{(1)} = 0$. On obtient les résultats suivants :



On a disposé les niveaux du test. On constate qu'un bas à gauche, qui correspond à l'hypothèse nulle, le seuil de 5% (qui est le seuil considéré dans ces simulations) est atteint (≈ 0.4978). De plus, on constate une augmentation de la puissance du test au fur-et-à-mesure de l'écartement à l'hypothèse nulle. Notamment, le test atteint une puissance proche de 1 pour les valeurs $\ell_n^{(1)} = \tau_n^{(1)} = 8$, ce qui est attendu. Les simulations illustrent donc le théorème 3 et 4.

Applications : étude des longitudes des taches solaires

Pour finir, nous avons appliqué notre test à la localisation des taches solaires. En effet, on dispose des localisations (longitudes et latitudes) des tâches solaires depuis la fin du 19ième siècle. Des études précédentes, on décrit les dépendances qui existent dans le processus des latitudes : la loi de Spörer décrit la dépendance sérielle (à l'aide d'un test de runs univarié) dans le processus des latitudes en valeurs absolues. Il y a de plus une indépendance du signe des latitudes. Autrement dit, les taches solaires se répartissent aléatoirement entre le nord et le sud mais avec une tendance à se rapprocher de l'équateur du Soleil au cours d'un cycle (qui dure aproximativement 11 ans).

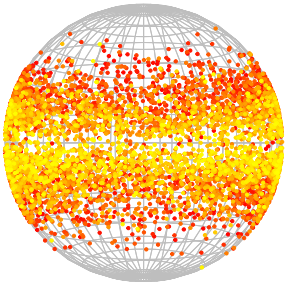


Figure 1: Données de localisations des tâches solaires au cours d'un cycle.

Pour les longitudes des tâches solaires, des précédents travaux ont mis en évidence une dépendance sérielle également dans les longitudes. Pour cela, une nouvelle statistique de runs directionnelles a été construite et étudiée. On renvoie à [1] pour plus de détails.

On s'est intéressé de savoir si le comportement des longitudes est le même au cours des différents cycles. Dans les données, nous disposons de $K = 14$ cycles : du cycle 11 au cycle 24.

Les tailles d'échantillons pour chaque cycle respectif sont :

$$251, 1916, 2860, 2291, 3527, 2938, 3586, 3593, 4664, 4708, 6195, 4551, 5373, 4850$$

En appliquant notre test, en choisissant $M = 5$, on obtient une p-valeur de 0.0001286 ce qui implique le rejet de l'hypothèse d'une loi commune entre les cycles pour les longitudes.

Bibliographie

[1] Maxime Boucher, Christian Francq, Yuichi Goto and Thomas Verdebout. On Runs Tests for Directional Data and Their Local and Asymptotic Optimality Properties (2025). Statistica Sinica.

[2] Eduardo García-Portugués and Pierre Lafaye de Micheaux and Simos G. Meintanis and Thomas Verdebout. Nonparametric tests of independence for circular data based on trigonometric moments (2021). Statistica Sinica.

[3] Jammalamadaka, S. R. and SenGupta, A. (2001). Topics in Circular Statistics, volume 5 of Series on Multivariate Analysis. World Scientific, Singapore.

[4] Meintanis, S. and Verdebout, T. (2019). Le Cam maximin tests for symmetry of circular data based on the characteristic function. Stat. Sin., 29(3):1301–1320.

[5] Ley, C. and Verdebout, T. (2017). Modern Directional Statistics. Chapman & Hall/CRC Interdis- ciplinary Statistics Series. CRC Press, Boca Raton.

Statistique mathématique pour données en grande dimension

Dependent default times with mixed hitting-time models

Mikael Escobar-Bach, Salima Helali [1], & Rejeb Hajer

[1] *Univ Paris-Est Créteil, LAMA, F-94010 Créteil, France,*
helali.salima@gmail.com

Résumé. Les modèles de temps d'atteinte mixte permettent d'analyser les risques concurrents à travers des décisions d'arrêt optimal, interprétées comme des temps de franchissement de processus de Lévy latents avec des seuils hétérogènes. Dans cet article, nous considérons un modèle bivarié de temps avec défaut dépendant, où les temps d'observation sont soumis à une censure et partagent un processus latent commun donné par un subordonnateur de Lévy. Nous établissons l'identifiabilité du modèle et proposons différents estimateurs pour les distributions marginales et la distribution de survie conjointe. Nous démontrons leurs propriétés asymptotiques et évaluons la performance en échantillon fini de nos résultats à travers une étude de simulation sur des données synthétiques, suivie d'une application utilisant des données réelles.

Mots-clés. Risque dépendant, modèle de temps d'atteinte mixte, processus de Lévy.

Abstract. Mixed hitting-time models allow the analysis of competing risks through optimal stopping decisions interpreted as crossing times of latent Lévy processes with heterogeneous thresholds. In this paper, we consider a bivariate time model with dependent default, where observation times are subject to censoring and share a common latent process given by a Lévy subordinator. We establish the identifiability of the model and propose different estimators for the marginal distributions and the joint survival distribution. We establish their asymptotic properties and evaluate the finite-sample performance of our results through a simulation study on synthetic data, followed by an application using real data.

Keywords. dependent risk, mixed hitting-time model, Lévy process.

1 Introduction

Threshold-crossing models for duration analysis have recently received considerable attention in various fields, including finance, reliability engineering and survival analysis. From a practical perspective, these models are especially useful in economics, with real applications such as patent races, technology adoption and smoking cessation among married couples in socio-economic settings contexts. The fundamental premise of the mixed hitting-time (MHT) models is to examine the time at which a process crosses a predetermined or random threshold, leading to significant events such as defaults, failures or other critical occurrences. This approach provides a flexible description of survival mechanisms (Aalen et al (2008)) in the context of optimal stopping time problems, where typical solutions are usually defined by threshold-crossing rules (Abbring (2012)). In this paper, we consider a bivariate competing risks model based on the threshold-crossing duration framework, where the two durations

correspond to the first-passage times of two dependent latent stochastic processes. The durations T and C are defined as the times when latent processes $(M_{1,t})_{t\geq 0}$ and $(M_{2,t})_{t\geq 0}$ exceed independent random barriers E_1 and E_2 respectively. This definition is often referred to as the *canonical construction* of default times (Bielecki and Rutkowski (2013)) and describes the standard intensity-based approach for modeling defaults. Specifically, we consider two proper cumulative hazard functions Λ_1 and Λ_2, that is, positive increasing functions on $\mathbb{R}_+$ such that $\Lambda_1(0) = \Lambda_2(0) = 0$. The default times are then defined by

$$T = \inf\{t \geq 0;\ M_{1,t} \geq E_1\}, \quad C = \inf\{t \geq 0;\ M_{2,t}(t) \geq E_2\} \tag{1.1}$$

and

$$M_{1,t} = \Lambda_1(t) + Y_t, \quad M_{2,t} = \Lambda_2(t) + Y_t \tag{1.2}$$

where $(Y_t)_{t\geq 0}$ is a right-continuous and increasing Lévy process, also referred to as a subordinator. In survival analysis, this issue is referred to as dependent censoring and occurs when only a lower bound of the event of interest is returned (Tsiatis (1975)).

2 Model and notations

Each such process $(Y_t)_{t\geq 0}$ is characterized by its marginal Lévy exponent κ, which is related to its Laplace-Stieltjes transformation ψ_{Y_t} by $\psi_{Y_t}(x) = e^{-t\kappa(x)}, t \geq 0$. In this work, we consider that the lifetimes of interest defined by the random times T and C are affected by the defaults of a Lévy process as defined in (1.1) where E_1 and E_2 are independent unit exponential random variables. In the sequel, we will also use the following equivalent formulation

$$T = \inf\{t \geq 0;\ e^{-M_{1,t}} \leq U_1\} \quad \text{and} \quad C = \inf\{t \geq 0;\ e^{-M_{2,t}} \leq U_2\},$$

where $U_1 = \exp(-E_1)$ and $U_2 = \exp(-E_2)$. Note that U_1 and U_2 are uniformly distributed. Based on a right-censoring approach, only the smallest default time is observed. Formally, the outcomes are limited to the couple (Z, δ), defined by

$$Z = \min(T, C) \quad \text{and} \quad \delta = \mathbb{1}_{\{T\leq C\}}.$$

3 Copulas and dependence properties

Proposition 1. *The joint survival function $\widetilde{P}$ of (T, C) defined in (1.1) is given, for any $t, s \geq 0$, by*

$$\widetilde{P}(t, s) = \overline{F}_C(s)\overline{F}_T(t) \min\left(\overline{F}_T(t)^{-\alpha(t)}, \overline{F}_C(s)^{-\beta(s)}\right),$$

where

$$\alpha(t) = \frac{t\xi}{\Lambda_1(t) + t\kappa(1)}, \qquad \beta(s) = \frac{\xi s}{\Lambda_2(s) + s\kappa(1)},$$

and $\xi = 2\kappa(1) - \kappa(2)$. In particular, the survival copula $\widetilde{\mathcal{C}}$ is given by

$$\widetilde{\mathcal{C}}(u, v) = uv \min\left(u^{-\alpha\left(\overline{F}_T^-(u)\right)}, v^{-\beta\left(\overline{F}_C^-(v)\right)}\right), \quad (u, v) \in \overline{F}_T(\mathbb{R}_+) \times \overline{F}_C(\mathbb{R}_+). \tag{3.1}$$

Proposition 2. *The survival function $\overline{H}$ of Z is given by $\overline{H}(t) = e^{-\Lambda_1(t)-\Lambda_2(t)-t\kappa(2)}$ for $t \geq 0$. Moreover, the sub-densities $f_{Z,\delta=0}$ and $f_{Z,\delta=1}$ of (Z,δ) are expressed by*

$$f_{Z,\delta=0}(t) = \overline{H}(t)[\Lambda_2'(t) + \kappa(2) - \kappa(1)] \quad \text{and} \quad f_{Z,\delta=1}(t) = \overline{H}(t)\left[\Lambda_1'(t) + \kappa(1)\right]. \tag{3.2}$$

In particular, the hazard functions λ_T and λ_C respectively associated with the random variables T and C are given by $\lambda_T = f_{Z,\delta=1}/\overline{H}$ and $\lambda_C = f_{Z,\delta=0}/\overline{H} + \xi$.

Proposition 3. *The model defined by (1.1)-(1.2) is a special case of the GMO model. Specifically, we have*

$$(T, C) \overset{d}{=} (\min(X_1, X_3), \min(X_2, X_3))$$

where $\overset{d}{=}$ denotes equality in distribution, X_1, X_2 and X_3 are three independent random variables such that X_3 follows an exponential distribution with parameter ξ. Moreover, the cumulative hazard functions $\Lambda_{X_1}, \Lambda_{X_2}$ of X_1 and X_2, respectively, are given by:

$$\begin{cases} \Lambda_{X_1}(t) = & \Lambda_1(t) + t\kappa(1) - \xi t \\ \Lambda_{X_2}(t) = & \Lambda_2(t) + t\kappa(1) - \xi t. \end{cases}$$

Remark 1. *Based on Proposition 2 in Escobar-Bach and Helali (2024), if Λ_1 and Λ_2 are strictly increasing, the Kendall's tau corresponding to the proposed model is given by*

$$\tau = 2\xi \int_0^{+\infty} \overline{H}(u)^2 du.$$

4 Identifiability

Proposition 4. *Assuming that $\Lambda_1'(0) = \Lambda_2'(0) = 0$, the model defined by (1.1)–(1.2) is identifiable.*

5 Estimation

We assume that we have an independent and identically distributed (*i.i.d.*) sample $(Z_i, \delta_i)_{1 \leq i \leq n}$ drawn from the observed model (Z, δ).

5.1 Non-parametric estimation of $\overline{F}_T$

According to Equation (3.2) in Proposition 2, it turns out that the Nelson-Aalen estimator $\Lambda_{T,n}$ of the cumulative hazard function Λ_T remains consistent, unlike most dependent censoring models. The Nelson-Aalen estimator of Λ_T is then defined by

$$\Lambda_{T,n}(t) = \sum_{i:Z_{(i)} \leq t} \frac{d_i}{n - i + 1} \quad \text{with} \quad d_i = \sum_{j=1}^{n} \mathbb{1}_{\{Z_j = Z_{(i)}, \delta_j = 1\}} \quad \text{and} \quad t > 0, \tag{5.1}$$

where $Z_{(i)}$ is the i-th order statistic among the n survival time observations. According to Lemma 1 of Breslow and Crowley (1974), we also conclude that the Kaplan-Meier (KM) estimator $\overline{F}_{T,n}$ defined by $\overline{F}_{T,n} = \exp(-\Lambda_{T,n})$ provides a consistent estimator of the survival function $\overline{F}_T$.

5.2 Semi-parametric estimation of $\overline{F}_C$, $\widetilde{P}$ and τ

Since the survival estimation $\overline{F}_T$ can be estimated by the KM estimator, we need to approximate the survival function $\overline{F}_C$ and the parameter ξ in order to estimate the survival joint distribution. To this end, we consider a semi-parametric approach based on maximum likelihood (ML) with Bernstein polynomial functions. Bernstein polynomials have been used by many authors in smoothing estimators and using functions of the form

$$s \in [0,1] \to \sum_{k=0}^{m} \phi_k b_{k,m}(s), \quad m \in \mathbb{N}$$

where m is the degree, $b_{k,m}(s) = C_m^k s^k (1-s)^{m-k}$ and $\overline{\phi}_m = (\phi_0, \ldots, \phi_m)^T \in \mathbb{R}^{m+1}$. In this context, we assume that both random times T and C have equal supports. For compact supports $S_{F_C} = S_{F_T} = [a,b]$, we can transform the data linearly to the interval $[0,1]$, using

$$\widetilde{Z} = (Z-a)/(b-a) = \min\left((T-a)/(b-a), (C-a)/(b-a)\right) = \min(\widetilde{T}, \widetilde{C}).$$

If $F_{\widetilde{C},m}$ and $g_{\widetilde{C},m}$ denote estimates of the distribution and the density functions of $\widetilde{C}$, respectively, then estimators of the function F_C and f_C can be obtained using the following relations:

$$F_{C,m}(t) = F_{\widetilde{C},m}\left(\frac{t-a}{b-a}\right) \quad \text{and} \quad f_{C,m}(t) = \frac{1}{b-a} f_{\widetilde{C},m}\left(\frac{t-a}{b-a}\right),$$

for $t \in [a,b]$. Transformations such as $\widetilde{Z} = Z/(1+Z)$ and $\widetilde{Z} = (1/2) + (\pi/2)\tan^{-1}(Z)$ can be used also to handle the cases of random variables with support $\mathbb{R}_+$ and $\mathbb{R}$, respectively, with the proper change of variables to recover $F_{C,m}$ and $f_{C,m}$. In the following, we consider that $S_{F_C} = S_{F_T} = [0,1]$. Based on Proposition 2, the log-likelihood is given by

$$\begin{aligned}
\mathcal{L}(\xi, \overline{F}_C, f_C) &= \sum_{i=1}^{n} \delta_i \log\{f_{Z,\delta=1}(Z_i)\} + (1-\delta_i)\log\{f_{Z,\delta=0}(Z_i)\} \\
&= \sum_{i=1}^{n} \delta_i \log\{f_Z(Z_i) - f_{Z,\delta=0}(Z_i)\} + (1-\delta_i)\log\{f_{Z,\delta=0}(Z_i)\},
\end{aligned}$$

where $f_{Z,\delta=0}(t) = \overline{F}_T(t) e^{t\xi}\left[f_C(t) - \xi \overline{F}_C(t)\right]$ and f_Z is the density of Z. The direct maximization of this likelihood can be challenging, since it depends on an unknown functions f_Z and $\overline{F}_T$. The idea is now to consider an approximation of the log-likelihood $\mathcal{L}$ through the following steps:

(i) we approximate f_Z using a non-parametric kernel method

$$f_{Z,n}(t) \;=\; \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t-Z_i}{h_n}\right), \tag{5.2}$$

where K represents a kernel function and $h_n > 0$ is a bandwidth parameter.

(ii) the function $\overline{F}_T$ is estimated using the KM estimator $\overline{F}_{T,n}$.

(iii) the functions F_C and f_C are approximated using the Bernstein polynomials. The estimators for the distribution function and the density function are then given by

$$F_{C,m}(t|\overline{\phi}_m) = \sum_{k=0}^{m} \phi_k b_{k,m}(t) \quad \text{and} \quad f_{C,m}(t) \;=\; \sum_{k=0}^{m} \phi_k b'_{k,m}(t), \tag{5.3}$$

where the coefficients $\overline{\phi}_m$ are replaced by estimates obtained from the ML approach.

Therefore, we approximate and parameterize the distribution F_C using a mixture of beta distributions, and we estimate $\overline{\phi}_m$ as parameters of the log-likelihood function $\mathcal{L}$. In summary, let $\mathcal{L}_m(\xi, \overline{\phi}_m)$ denote the approximate log-likelihood obtained by replacing $\overline{F}_T$, f_Z, $\overline{F}_C$ and f_C with $\overline{F}_{T,n}$, $f_{Z,n}$, $\overline{F}_{C,m}$ and $f_{C,m}$ respectively. We consider the parameter vector $\theta = (\overline{\phi}_m, \xi)$ and define the ML estimator

$$\theta_n = \underset{\theta\in\mathbb{R}^{m+2}}{\operatorname{argmax}} \mathcal{L}_m(\theta).$$

Based on the estimator $\theta_n = \left(\overline{\phi}_{n,m}, \xi_n\right)$, we obtain the estimators of the survival function $\overline{F}_C$ and the joint law $\widetilde{P}$ namely

$$\overline{F}_{C,m,n}(t) = 1 - \sum_{k=0}^{m} \phi_{k,n} b_{k,m}(t),$$

$$\widetilde{P}_{m,n}(t,s) = \overline{F}_{C,m,n}(s)\overline{F}_{T,n}(t) \min\left(\overline{F}_{T,n}(t)^{-\alpha_n(t)}, \overline{F}_{C,m,n}(s)^{-\beta_{m,n}(s)}\right),$$

where

$$\alpha_n(t) = \frac{\xi_n t}{-\ln(\overline{F}_{T,n}(t))} \qquad \text{and} \qquad \beta_{m,n}(t) = \frac{\xi_n t}{-\ln(\overline{F}_{C,m,n}(t))}.$$

5.3 Asymptotic properties

We consider the optimization problem based on the minimization of the Kullback-Leibler information criterion given by $I(g,h;\theta) = \mathbb{E}\left(\log\left[g(Z,\delta;\xi)/h(Z,\delta;\theta)\right]\right)$ with densities

$$g(t,b;\xi) = \left[f_Z(t) - \overline{F}_T(t)e^{t\xi}\left(f_C(t) - \xi\overline{F}_C(t)\right)\right]^b \left[\overline{F}_T(t)e^{t\xi}\left(f_C(t) - \xi\overline{F}_C(t)\right)\right]^{1-b},$$

and

$$\begin{aligned} & h(t,b;\theta) \\ = \; & \left[f_{Z,n}(t) - \overline{F}_{T,n}(t)e^{t\xi}\left(f_{C,m,n}(t|\overline{\phi}_m) - \xi\overline{F}_{C,m,n}(t|\overline{\phi}_m)\right)\right]^b \left[\overline{F}_{T,n}(t)e^{t\xi}\left(f_{C,m,n}(t|\overline{\phi}_m) - \xi\overline{F}_{C,m,n}(t|\overline{\phi}_m)\right)\right]^{1-b}, \end{aligned}$$

with $\overline{F}_{T,n}$, $f_{Z,n}$, $\overline{F}_{C,m,n}$ and $f_{C,m,n}$ the estimators defined in the previous section. Let us introduce the following matrices

$$A(\theta) = \left(\mathbb{E}\left[\frac{\partial^2 \log h(Z,\delta;\theta)}{\partial\theta_i \partial\theta_j}\right]\right)_{1\le i,j\le m+2} \text{ and } \qquad B(\theta) = \left(\mathbb{E}\left[\frac{\partial \log h(Z,\delta;\theta)}{\partial\theta_i}\frac{\partial \log h(Z,\delta;\theta)}{\partial\theta_j}\right]\right)_{1\le i,j\le m+2}.$$

($\mathbf{H_1}$): $\Lambda_2' e^{-\Lambda_2}$ and $\Lambda_1' e^{-\Lambda_1}$ are element of $L^2(0,\infty)$.

($\mathbf{H_2}$): The kernel K is bounded up by a positive constant K_{max}.

($\mathbf{H_3}$): I has unique minimum at $\theta^\star$.

($\mathbf{H_4}$): (a) $\theta^\star$ is interior to $\mathbb{R}_+^{m+2}$. (b) $B(\theta^\star)$ is non singular. (c) $\theta^\star$ is a regular point of $A(\theta)$.

Proposition 5. *Suppose that Λ_1 and Λ_2 satisfy ($\mathbf{H_1}$). Then, all moments of order k for Z exist, that is*

$$\mathbb{E}(Z^k) < \infty.$$

Proposition 6. *Under assumptions ($\mathbf{H_1}$) – ($\mathbf{H_4}$), θ is identifiable and*

$$\sqrt{n}\left(\theta_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, C\left(\theta\right)\right),$$

where $C(\theta) = A(\theta)^{-1} B(\theta) A(\theta)^{-1}$.

Proposition 7. *Under assumptions ($\mathbf{H_1}$) – ($\mathbf{H_4}$),*

$$\lim_{n,m\to\infty} \widetilde{P}_{n,m}(t,s) = \widetilde{P}(t,s) \text{ and } \lim_{n\to\infty} \tau_n = \tau.$$

References

Aalen, O., Borgan, O., and Gjessing, H., (2008). Survival and event history analysis: a process point of view. *Springer Science and Business Media.*

Abbring, J. H., (2012). Mixed Hitting-Time Models. *Econometrica*, **80** 783–819.

Bielecki, T. R., and Rutkowski, M., (2013). Credit risk: modeling, valuation and hedging. *Springer Science and Business Media.*

Breslow, N. and Crowley, J., (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of statistics*, **2**, 437–453.

Escobar-Bach, M., and Helali, S., (2024). Dependent censoring with simultaneous death times based on the Generalized Marshall-Olkin model. *Journal of Multivariate Analysis,* **105347.**

Tsiatis A., (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, **72**, 20–22.

Calibration data-dépendante pour un test à noyau non-asymptotique à deux échantillons

Perrine Lacroix [1] & Bertrand Michel [2] & Franck Picard [3] & Vincent Rivoirard [4]

[1] *Nantes Université, Laboratoire de mathématiques Jean Leray, F-44000, Nantes.*
[2] *Nantes Université, Centrale Nantes, Laboratoire de mathématiques Jean Leray, F-44000, Nantes.*
[3] *Laboratoire de Biologie et Modélisation de la Cellule, École Normale Supérieure de Lyon, CNRS, UMR5239, Université Claude Bernard Lyon 1, Lyon.*
[4] *Ceremade, CNRS, UMR 7534, Université Paris-Dauphine, PSL University, Paris.*

perrine.lacroix@univ-nantes.fr ; Bertrand.Michel@ec-nantes.fr ;
franck.picard@ens-lyon.fr ; Vincent.Rivoirard@dauphine.fr

Résumé. Soient deux populations de données multivariées décrites par p variables, où p est significativement plus grand que les tailles des populations. Nous souhaitons déterminer, via un test statistique à deux échantillons, si les distributions dont sont issues les deux populations sont égales (hypothèse nulle $H_0 = \{\mathbb{P}_1 = \mathbb{P}_2\}$) ou différentes (hypothèse alternative $H_1 = \{\mathbb{P}_1 \neq \mathbb{P}_2\}$). Pour tenir compte du critère multivarié, de grande dimension et tenir compte des fortes dépendances entre les variables, nous proposons un test de comparaison des deux distributions basé sur les méthodes à noyaux : nos données sont au préalable transformées via une fonction de plongement bien choisie et vivent dans un espace de Hilbert à noyau reproduisant (RKHS). Nous proposons une statistique de test inspirée de Harchaoui et al. (2008) généralisant le test de Student (ou le test de T^2 de Hotelling) dans un RKHS. Ici, la loi de notre statistique de test, qui n'est pas explicite, est inconnue sous l'hypothèse nulle H_0 et l'opérateur de covariance est régularisé par troncature de sa décomposition spectrale. Nous proposons une estimation du quantile à la taille d'échantillon fixée, ceci via l'obtention d'une borne théorique non-asymptotique de notre statistique de test. Sur le plan applicatif, nous proposons un algorithme qui garantit un contrôle de l'erreur de type I (se tromper sous H_0). Cet algorithme repose uniquement sur les données disponibles et inclut la calibration de l'hyperparamètre de troncature.

Mots-clés. tests statistiques, méthodes à noyau, non-asymptotique, calibration data-dépendante, comparaison de populations de cellules.

Abstract. We consider two populations of multivariate data described by p variables, where p is significantly larger than the population sizes. A two-sample statistical test is performed to decide whether the distributions from which the two populations are derived are equal (null hypothesis $H_0 = \{\mathbb{P}_1 = \mathbb{P}_2\}$) or different (alternative hypothesis $H_1 = \{\mathbb{P}_1 \neq \mathbb{P}_2\}$). To consider the multivariate, the high-dimensional context and the strong dependencies between the variables, we propose a comparison test based on kernel methods : our data are first embedded in a well-chosen Reproducing Kernel Hilbert Space

(RKHS). We propose a test statistic inspired by Harchaoui et al (2008), generalizing the Student t-test (or Hotelling's T^2 test) into an RKHS. Here, the distribution of our test statistic, which is not explicit, is unknown under the null hypothesis H_0 and the covariance operator is regularized by its truncated spectral decomposition. We propose an estimator of the quantile for a fixed sample size, through a non-asymptotic theoretical bound on our test statistic. For applications, we propose an algorithm that guarantees type I error control (being wrong under H_0). This algorithm only uses the available data and includes calibration of the truncation hyperparameter.

Keywords. statistical tests, kernel methods, non-asymptotic, data-dependent calibration, comparison of cell populations.

1 Nouveaux défis statistiques pour la génomique

Grâce aux nouvelles technologies de pointe en génomique, les nouveaux jeux de données contiennent de l'information simultanée à la fois sur tous les gènes et à la fois sur toutes les cellules d'un individu. Ces changements d'échelle permettent entre autres (i) d'étudier les liens entre les cellules [Efremova et al., 2020] et de mettre en évidence des types cellulaires rares [Andrews and Hemberg, 2018] ; (ii) de comprendre l'hétérogénéité et la variabilité biologique d'une cellule au sein d'un micro-environnement ; (iii) d'avoir accès à la trajectoire (discrète) du gène le long des cellules de l'organisme, et de prendre en compte la dépendance entre les gènes. En contrepartie, la taille des jeux de données a explosé et la complexité des structures présentes dans ces jeux a augmenté considérablement. Pour les données décrivant l'activité génomique de chaque cellule d'un organisme (données transcriptomiques de séquençage à haut débit "scRNA-Seq"), le nombre de cellules peut atteindre $100,000$ pour un nombre de gènes autour de $30,000$. Prendre en compte de tels jeux de données dans les analyses statistiques d'aujourd'hui est un réel challenge à relever.

Ce travail a été motivé par la comparaison de deux groupes de données génomiques. C'est un problème qui concerne de nombreuses applications. On peut citer l'identification de populations cellulaires, la caractérisation de la différenciation cellulaire en réponse à des changements environnementaux, la comparaison d'une cellule à un référent sain ou pathologique, etc. La prise en compte de ces jeux de données massifs, complexes et riches en information a été réalisée dans un cadre asymptotique [Ozier-Lafontaine et al., 2023]. C'est une alternative non-asymptotique que nous proposons ici.

2 Tests à noyau à deux échantillons

Soient les deux populations de données $(x_1, \cdots, x_{n_1})$ et $(y_1, \cdots, y_{n_2})$ dont nous supposons être un n_1-échantillon (resp. n_2-échantillon) issu d'une distribution $\mathbb{P}_1$ (resp. $\mathbb{P}_2$). D'un point de vue statistique, dire que les deux conditions biologiques partagent les

mêmes caractéristiques est équivalent à dire que les deux distributions $\mathbb{P}_1$ et $\mathbb{P}_2$ sont égales. Ainsi, nous nous intéressons aux tests statistiques non paramétriques multivariés à deux échantillons pour évaluer l'hypothèse nulle $H_0 = \{\mathbb{P}_1 = \mathbb{P}_2\}$ contre son alternative $H_1 = \{\mathbb{P}_1 \neq \mathbb{P}_2\}$.

2.1 Tests basés sur les méthodes à noyau

Pour capturer les relations complexes entre les variables et éliminer certains problèmes issus de la grande dimension, les données sont au préalable envoyées dans un espace de Hilbert à noyau reproduisant (RKHS), noté $\mathcal{H}$, via une fonction de plongement ϕ judicieusement choisie. Cet espace, plus complexe, permet l'utilisation des méthodes linéaires [Muandet et al., 2017] réduisant considérablement la complexité du problème (figure 1).

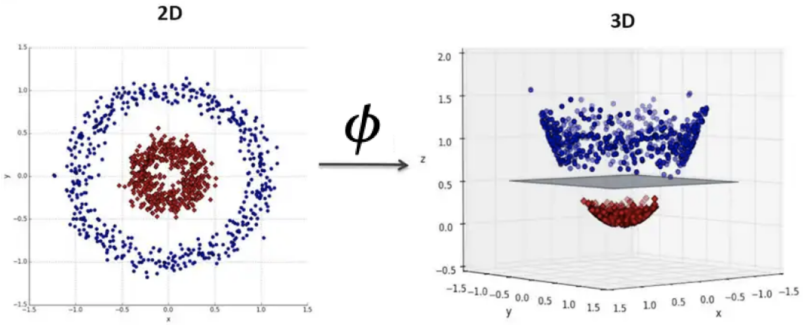


FIGURE 1 – Illustration du plongement des données dans un RKHS : la séparation des deux populations devient linéaire dans le RKHS (ici $\mathbb{R}^3$).

Pour une fonction ϕ dite *caractéristique* (ce que nous supposons par la suite), le test statistique $H_0 = \{\mathbb{P}_1 = \mathbb{P}_2\}$ contre $H_1 = \{\mathbb{P}_1 \neq \mathbb{P}_2\}$ est équivalent au test fonctionnel sur les moyennes dans $\mathcal{H}$:

$$H_0 = \{\mu_1 = \mu_2\} \quad \text{et} \quad H_1 = \{\mu_1 \neq \mu_2\} \tag{1}$$

où μ_1 et μ_2 sont les moyennes des données plongées dans $\mathcal{H}$. Elles caractérisent ainsi entièrement dans $\mathcal{H}$ les distributions des deux populations, ce qui fait des méthodes à noyau un outil extrêmement flexible et puissant (voir figure 1.1 de [Muandet et al., 2017]).

2.2 Statistique de test de T^2 de Hotelling à noyau avec troncature spectrale

La statistique naturelle pour le test (1) est le *Maximum Mean Discrepancy (MMD)* [Gretton et al., 2012]) :

$$\widehat{MMD} = \|\widehat{\mu}_1 - \widehat{\mu}_2\|^2_{\mathcal{H}},$$

où $\widehat{\mu}_1$ et $\widehat{\mu}_2$ sont les moyennes empiriques des données dans $\mathcal{H}$ des deux distributions. Mais la puissance associée à cette statistique n'est pas très élevée due à la non prise en compte de la variabilité intra-population.

Une alternative est *la statistique de test de T^2 de Hotelling à noyau* [Harchaoui et al., 2008] :

$$\widehat{D}^2 = \frac{n_1 n_2}{n_1 + n_2} \left\| \widehat{\Sigma}_W^{-\frac{1}{2}} \left(\widehat{\mu}_1 - \widehat{\mu}_2 \right) \right\|_{\mathcal{H}}^2$$

où $\widehat{\Sigma}_W$ est l'opérateur de variance-covariance défini par :

$$\widehat{\Sigma}_W = \frac{n_1}{n_1 + n_2} \left(\frac{1}{n_1} \sum_{i=1}^{n_1} \left(\phi(x_i) - \widehat{\mu}_1 \right)^{\otimes^2_{\mathcal{H}}} \right) + \frac{n_2}{n_1 + n_2} \left(\frac{1}{n_2} \sum_{j=1}^{n_2} \left(\phi(y_j) - \widehat{\mu}_2 \right)^{\otimes^2_{\mathcal{H}}} \right). \quad (2)$$

Mais l'opérateur de covariance $\widehat{\Sigma}_W$ est mal conditionné (il a au plus $n_1 + n_2$ valeurs propres non nulles), il n'est pas inversible et (2) est mal définie.

Nous proposons de régulariser l'opérateur $\widehat{\Sigma}_W$ par troncature de sa décomposition spectrale, plus adaptée en pratique (pour l'interprétabilité notamment) que la régularisation classique Ridge (déjà étudiée dans [Hagrass et al., 2023]). La statistique sous-jacente est :

$$\widehat{D}_T^2 = \frac{n_1 n_2}{n_1 + n_2} \sum_{t=1}^{T} \widehat{\lambda}_t^{-\frac{1}{2}} \langle \widehat{f}_t, \widehat{\mu}_1 - \widehat{\mu}_2 \rangle_{\mathcal{H}} \quad (3)$$

où $\widehat{\lambda}_t, \widehat{f}_t$ sont les t-ièmes éléments de la décomposition spectrale de $\widehat{\Sigma}_W$ et T est le paramètre de troncature inconnu.

3 Borne théorique et calibration des quantiles de la statistique de test (3)

3.1 Inégalité de concentration pour la statistique de test (3)

Puisque $\widehat{D}_T^2$ est proche de 0 sous H_0 et s'en éloigne sous H_1, l'objet d'intérêt pour calibrer le test est le quantile $q_{1-\alpha}$ tel que l'erreur de première espèce $\mathbb{P}_{H_0}\left(\widehat{D}_T^2 > q_{1-\alpha} \right)$ soit contrôlée par un niveau $\alpha \in]0, 1[$ fixé. La loi de $\widehat{D}_T^2$ sous H_0 étant inconnue (et donc le quantile indisponible), deux alternatives existent à T fixé : approcher le quantile asymptotiquement par le quantile de la loi du Khi-deux de paramètre T [Ozier-Lafontaine et al., 2023] ou approcher la loi sous H_0 via des techniques d'agrégation ou de permutations de tests. Ces deux approches ne sont cependant pas adaptées à la grande dimension ou sont coûteuses en temps de calcul et en mémoire de stockage.

Dans cet exposé, je présenterai une nouvelle approche pour calibrer le test en proposant une estimation du quantile à la taille d'échantillon fixée. Celle-ci est obtenue via un contrôle théorique et non-asymptotique des quantiles de (3). Plus précisément, dans le cadre équilibré ($n_1 = n_2 = n$) et sous certaines hypothèses spectrales, nous avons montré l'existence de constantes absolues C_1, C_2, C_3 telles que pour tout $\delta > 0$:

$$\mathbb{P}_{H_0}\left(\widehat{D}_T^2 > 8T\delta \max_{t=1,\cdots,T}\left(\frac{\lambda_t - C_1\cdot\sqrt{\frac{\delta}{n}}}{\lambda_t - \sqrt{\frac{\delta}{n}}\left(\frac{C_2}{\tilde{\Delta}_t} + C_3\right)}\right)\right) \leq 9Te^{-\delta} \tag{4}$$

où $\tilde{\Delta}_t := \delta_t \vee \delta_{t-1}$ avec $\delta_t := \frac{1}{2}(\lambda_t - \lambda_{t+1})$.

3.2 Calibration data-dépendante du test

Les constantes dans (4) sont explicites mais très grandes. L'utilisation brute d'un tel quantile $x := 8T\delta \max_{t=1,\cdots,T}\left(\frac{\lambda_t - C_1\cdot\sqrt{\frac{\delta}{n}}}{\lambda_t - \sqrt{\frac{\delta}{n}}\left(\frac{C_2}{\tilde{\Delta}_t} + C_3\right)}\right)$ conduirait à une conservation systématique de l'hypothèse H_0. Ainsi, pour les visées applicatives, nous proposons un algorithme basé sur (4) mais où les constantes C_1, C_2 et C_3 ainsi que le paramètre de troncature T sont considérés comme des hyperparamètres inconnus qui sont calibrés directement à partir du jeu de données disponible. Cet algorithme permet de garantir un contrôle (expérimental pour le moment) de l'erreur de type I (probabilité de se tromper sous H_0) (figure 2).

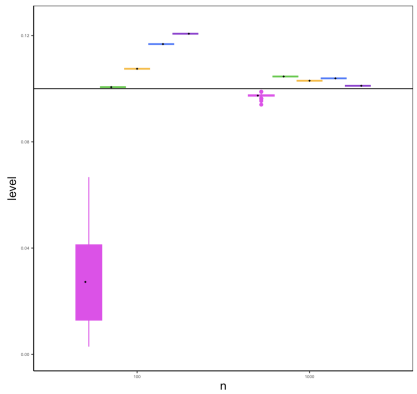


FIGURE 2 – Erreurs de type I obtenues via le quantile x (rose) et les quantiles de la loi du Khi-deux de paramètres 1 (vert), 2 (jaune), 3 (bleu) et 4 (violet) pour $n = 100$ et 1000 et $p = 1$. Ici, $\mathbb{P}_0 = \mathbb{P}_1 = \mathcal{N}(0,1)$, le noyau est gaussien et $\alpha = 0.1$. L'erreur de type I est approchée par : $\frac{1}{M}\sum_{m=1}^{M}\frac{1}{R}\sum_{r=1}^{R}1_{\left\{\widehat{D}_T^2(r) > x_{1-\alpha,m}\right\}}$, où $R = 10.000$ et $M = 1.000$, $\widehat{D}_T^2(r)$ est calculée sur le r-ième jeu de données et $x_{1-\alpha,m}$ est calculé sur le m-ième jeu de données.

Références

[Andrews and Hemberg, 2018] Andrews, T. S. and Hemberg, M. (2018). Identifying cell populations with scrnaseq. *Molecular aspects of medicine*, 59 :114–122.

[Efremova et al., 2020] Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. (2020). Cellphonedb : inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature protocols*, 15(4) :1484–1506.

[Gretton et al., 2012] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1) :723–773.

[Hagrass et al., 2023] Hagrass, O., Sriperumbudur, B. K., and Li, B. (2023). Spectral regularized kernel goodness-of-fit tests. *arXiv preprint arXiv :2308.04561*.

[Harchaoui et al., 2008] Harchaoui, Z., Bach, F., and Moulines, E. (2008). Testing for Homogeneity with Kernel Fisher Discriminant Analysis. *arXiv :0804.1026 [stat]*.

[Muandet et al., 2017] Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. (2017). Kernel mean embedding of distributions : A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2) :1–141.

[Ozier-Lafontaine et al., 2023] Ozier-Lafontaine, A., Fourneaux, C., Durif, G., Vallot, C., Gandrillon, O., Giraud, S., Michel, B., and Picard, F. (2023). Kernel-based testing for single-cell differential analysis.

Asymptotic subsampling for supervised learning with reproducing kernel Hilbert spaces

Maxime Sangnier[1] & Eyal Vayness[2]

[1,2]*Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France*
[1]*maxime.sangnier@sorbonne-universite.fr* [2]*eyal.vayness@sorbonne-universite.fr*

Résumé. Le sous-échantillonnage est une méthode statistique dont le but est de réduire le coût computationel induit par l'accumulation de grands jeux de données, notamment à l'ère du Big data. En pratique, cela permettrait de réduire les émissions carbone et la consommation énergétique. Dans un plan de sous-échantillonnage avec remise il apparait une variante du risque empirique usuel repondéré par l'inverse des probabilités, qui est un estimateur d'Horvitz-Thompson de ce dernier. Les outils existants de l'echantillonnage préférentiel donnent des résultats sur les probabilités de sous-échantillonnage pour que le risque sous-échantillonné soit le plus proche possible de son pendant empirique. Notre méthode mobilise les propriétés asymptotiques du minimiseur du risque sous-échantillonné (MRS), par rapport au minimiseur du risque empirique, pour extraire une loi de sous-échantillonnage (LSE) optimale. Nous montrons d'abord la normalité asymptotique du MRS, avec un opérateur de covariance qui dépend de la LSE, et proposons ensuite une stratégie de sous-échantillonage qui est L-optimale.

Mots-clés. Sous-échantillonnage, Asymptotique, RKHS, Dispositif expérimental optimal, L-optimalité.

Abstract. Subsampling is a statistical method, the goal of which is to reduce the computational toll that is induced by the aggregation of large datasets, especially in the era of Big data. In practice, it may lead to carbon emissions and energy consumption reduction. In a subsampling with replacement scheme one deals with an inverse probability weighted version of the usual empirical objective, that is a Horvitz-Thompson-like estimator of the latter. Existing tools from importance sampling theory yield results on the subsampling probabilities (SSP) so that the subsampled objective is as close as possible to the empirical one. We leverage the asymptotic properties of the subsampled risk minimizer (SRM), with respect to the empirical risk minimizer (ERM), in order to extract another optimal SSP. We first show that the SRM is asymptotically normal around the ERM with a covariance operator that depends on the subsampling scheme, and then propose an SSP that is L-optimal.

Keywords. Subsampling, Asymptotic, RKHS, Optimal experimental design, L-optimality.

1 Introduction

As of the last couple decades, the disproportionate data accumulation in various domains such as medical research, road safety, bioinformatics, physics, genetics and social media, led scientists to rethink standard statistical learning methods. Beyond the computational barrier, in terms of time and memory, and the economic cost of training models on large datasets, the overconsumption of energy resources and ethic excesses also became concerning topics. Limiting data usage may be done at estimation time, for instance with a stochastic version of a gradient descent algorithm (see Bottou, Curtis, and Nocedal, 2018 for a review of stochastic gradient descent methods), or beforehand by meticulously selecting a subset of the available data, as it is done in subsampling. Getting rid of pieces of information lighten the estimation phase at the expense of statistical richness.

Subsampling is a large domain and can be seen from different point of views including coresets (Braverman et al., 2021; Huggins, Campbell, and Broderick, 2016), importance sampling (Clemencon, Bertail, and Papa, 2016) or sketching (Ahmad et al., 2024) among others. Coresets and sketching are a mean to summarize available data. On the other hand, importance sampling aims at reducing the variance of the subsampled risk, as it is seen as an estimator of the empirical risk (itself being an estimator of the true risk). On the contrary, asymptotic subsampling leverages asymptotic properties of the subsampled risk minimizer (SRM) to derive an *optimal* subsampling scheme (see Wang, Zou, and Wang, 2022 for the theory in finite dimension).

In this communication, we set focus on studying procedures allowing to limit the volume of data to use during estimation of nonparametric models in statistical learning within the framework of reproducing kernel Hilbert spaces (RKHSs). We show usual asymptotic properties, namely consistency and asymptotic normality, of the empirical risk minimizer (ERM) (see (Sancetta, 2021) for the ERM's asymptotic behaviors on different assumptions), and of the SRM. We then propose a subsampling scheme that is L-optimal.

2 Mathematical settings

Let $\mathcal{D}_n = \{Z_1, \dots, Z_n\}$ be a dataset of i.i.d. copies of a generic random variable of interest Z of which we have a canonical decomposition $Z = (X, Y)$ so that the covariates X lies in $\mathcal{X} \subset \mathbb{R}^p$ and the response Y is in $\mathcal{Y} \subset \mathbb{R}$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous positive definite kernel and denote $\mathbb{H}$ its associated RKHS of functions from $\mathcal{X}$ to $\mathcal{Y}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ a convex and differentiable loss function associated to the problem. The true function of interest is h_0, defined as the minimizer of the true risk denoted $M : h \mapsto \mathbb{E}\left[\ell(h(X), Y)\right]$. Likewise, the ERM $\hat{h}_n$ is the minimizer of the empirical risk

$$M_n(h) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i). \tag{1}$$

Denote K_x to be the element $k(x, \cdot)$ of $\mathbb{H}$ for all $x \in \mathcal{X}$ and "$\otimes$" the outer product on the RKHS, i.e. for $h_1, h_2 \in \mathbb{H}$, $(h_1 \otimes h_2)(h) \overset{\text{def}}{=} \langle h_1, h \rangle_{\mathbb{H}} h_2$ for all $h \in \mathbb{H}$. Under standard

assumptions, the ERM is consistent and asymptotically normal. Simply, these assumptions are that the covariates and the predictors lie in bounded closed sub-spaces, the loss function ℓ is twice continuously differentiable and the underlying data distribution has finite fourth moment.

Theorem 2.1 (Informal)**.** *With some regularity conditions on the problem, one has consistency and asymptotic normality of the ERM, i.e.* $\|\hat{h}_n - h_0\|_\infty \xrightarrow[n\to\infty]{\mathbb{P}} 0$ *and*

$$\sqrt{n}\ddot{M}(\hat{h}_n - h_0) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{G}(0, \Sigma), \tag{2}$$

with $\ddot{M} \overset{\text{def}}{=} \mathbb{E}[\ell''(h_0(X), Y)K_X \otimes K_X]$ *and* $\Sigma \overset{\text{def}}{=} \mathbb{E}[\ell'(h_0(X), Y)^2 K_X \otimes K_X]$ *where* ℓ' *(resp.* ℓ''*) is the first (resp. second) partial derivative of the loss* ℓ *w.r.t. its first argument and* $\mathcal{G}(0, \Sigma)$ *is a zero-mean Gaussian process with covariance operator* Σ*.*

2.1 Subsampling with replacement

In a resampling with replacement framework, the subsampled risk is constructed as follows: given a probability vector $\boldsymbol{\pi}_n = (\pi_{n,1}, \ldots, \pi_{n,n})$, define the subsampled risk, for $s_n \in \mathbb{N}$ resampled indices, of an estimator $h \in \mathbb{H}$ as,

$$M^*_{s_n}(h; \boldsymbol{\pi}_n) \mapsto \frac{1}{s_n} \sum_{j=1}^{s} \frac{\ell(h(X_{I_j}), Y_{I_j})}{n\pi_{n,I_j}}, \tag{3}$$

where $I_1, \ldots, I_{s_n} \overset{i.i.d.}{\sim} \sum_{i=1}^n \pi_{n,i}\delta_{\{i\}}$ and denote $h^*_{s_n}$ its minimizer, which we call the SRM.

The first step towards asymptotic driven subsampling is to establish normality of the SRM, with an explicit sequence of covariance operator estimators. The regularity conditions for the below theorem are the same as the previous one, with the additional assumption that no subsampling probability goes to 0 faster than $1/n$.

Theorem 2.2 (Informal)**.** *With some regularity conditions on the problem, one has consistency and asymptotic normality of the SRM if* $s_n \xrightarrow[n\to\infty]{} \infty$*, i.e.* $\|h^*_{s_n} - \hat{h}_n\|_\infty \xrightarrow[n\to\infty]{\mathbb{P}} 0$ *and*

$$\sqrt{s_n} V_n^{-1/2}(h^*_{s_n} - \hat{h}_n) \xrightarrow[n\to\infty]{\mathrm{d}} \mathcal{G}(0, \mathrm{id}),$$

with $V_n = \ddot{M}^{-1}\Sigma_n(\boldsymbol{\pi})\ddot{M}^{-1}$ *with* $\Sigma_n(\boldsymbol{\pi}) = n^{-2}\sum_{i=1}^n \ell'(\hat{h}_n(X_i), Y_i)^2/\pi_{n,i} K_{X_i} \otimes K_{X_i}$*.*

It is common in asymptotic, and experimental designs to minimize a criterion based on the asymptotic covariance operator (Imberg, Axelson-Fisk, and Jonasson, 2023). We focus here on the L-optimality criterion, which minimizes the trace of a linear transformation L of the asymptotic covariance operator. It provides a simpler criterion to minimize for a well chosen L as one can get read of some Hessian computation and matrix inversion. Thus, if considering $L = \ddot{M}$, one aims to find the minimizer of

$$\boldsymbol{\pi}_n \mapsto \mathrm{Tr}(\ddot{M}\,\mathrm{Var}(h^*_{s_n})) = \mathrm{Tr}(\Sigma_n(\boldsymbol{\pi}_n)) = \frac{1}{n^2}\sum_{i=1}^n \frac{\ell'(\hat{h}_n(X_i), Y_i)^2}{\pi_{n,i}} k(X_i, X_i).$$

It follows that the probability vector $\boldsymbol{\pi}_n^\star$ given by

$$\pi_{n,i}^\star \overset{\text{def}}{=} \frac{|\ell'(\hat{h}_n(X_i), Y_i)|\sqrt{k(X_i, X_i)}}{\sum_{j=1}^n |\ell'(\hat{h}_n(X_j), Y_j)|\sqrt{k(X_j, X_j)}}, \tag{4}$$

for $i \in \{1, \ldots, n\}$, is the optimal subsampling probability (SSP) vector under L-optimality.

In practice, these optimal SSPs are undesirable as they call for the computation of the ERM $\hat{h}_n$. Therefore, the latter is replaced in Equation (4) by a *pilot* estimator $h_{s_0}^{\mathrm{P}}$. It is usually the minimizer of $M_{s_0}^*(\,\cdot\,; \boldsymbol{\pi}_n^{\mathrm{U}})$, for some $s_0 \ll s_n$, with $\boldsymbol{\pi}_n^{\mathrm{U}} = (1/n, \ldots, 1/n)$, the uniform probability vector. Moreover, to act as regularization, the resulting SSP vector $\boldsymbol{\pi}_n^{\mathrm{P}}$ is then transformed following a convex combination with $\boldsymbol{\pi}_n^{\mathrm{U}}$.

3 Discussions

In this communication, we exhibited a subsampling scheme that is asymptotically optimal. Our future work will focus on empirical comparison of this subsampling scheme alongside some baselines and other existing approximations methods, such as Nystroem (Williams and Seeger, 2000) and random Fourier features (Rahimi and Recht, 2007). As the premise of subsampling is to relieve the computational burden of using all available datapoints, a fair basis for comparison would be execution time of these algorithms. On the one hand, it is clear that the training time is reduced as less samples are utilized. On the other hand, we added a new step in the process, namely the SSP computation. This trade-off between execution time and accuracy could be an interesting direction for future investigation.

References

Ahmad, T. E., L. Brogat-Motte, P. Laforgue, and F. d'Alché-Buc (2024). "Sketch In, Sketch Out: Accelerating Both Learning and Inference for Structured Prediction with Kernels". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 109–117.

Bottou, L., F. E. Curtis, and J. Nocedal (2018). "Optimization Methods for Large-Scale Machine Learning". In: *SIAM Review* 60.2, pp. 223–311.

Braverman, V., D. Feldman, H. Lang, A. Statman, and S. Zhou (2021). "Efficient Coreset Constructions via Sensitivity Sampling". In: *Proceedings of The 13th Asian Conference on Machine Learning*. PMLR, pp. 948–963.

Clemencon, S., P. Bertail, and G. Papa (2016). "Learning from Survey Training Samples: Rate Bounds for Horvitz-Thompson Risk Minimizers". In: *Proceedings of The 8th Asian Conference on Machine Learning*. PMLR, pp. 142–157.

Huggins, J., T. Campbell, and T. Broderick (2016). "Coresets for Scalable Bayesian Logistic Regression". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc.

Imberg, H., M. Axelson-Fisk, and J. Jonasson (2023). *Optimal Subsampling Designs*. arXiv: 2304.03019 [math].

Rahimi, A. and B. Recht (2007). "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems*. Vol. 20. Curran Associates, Inc.

Sancetta, A. (2021). "Estimation in Reproducing Kernel Hilbert Spaces With Dependent Data". In: *IEEE Transactions on Information Theory* 67.3, pp. 1782–1795.

Wang, J., J. Zou, and H. Wang (2022). "Sampling With Replacement vs Poisson Sampling: A Comparative Study in Optimal Subsampling". In: *IEEE Transactions on Information Theory* 68.10, pp. 6605–6630.

Williams, C. and M. Seeger (2000). "Using the Nyström Method to Speed Up Kernel Machines". In: *Advances in Neural Information Processing Systems*. Vol. 13. MIT Press.

Statistique multivariée

Propriétés et robustesse de la profondeur scatter

Gaëtan Louvet [1] & Germain Van Bever [2] & Stanislav Nagy [3]

[1] *Université libre de Bruxelles, Belgique, gaetan.louvet@ulb.be*
[2] *Université libre de Bruxelles, Belgique, germain.van.bever@ulb.be*
[3] *Charles University, Czech Republic, nagy@karlin.mff.cuni.cz*

Résumé. La notion de profondeur statistique fournit des outils nonparamétriques robustes pour analyser les distributions. Les fonctions de profondeur mesurent en effet l'adéquation entre des paramètres et les lois de probabilité sous-jacentes. Dans le cas *position*, la profondeur de demi-espace a déjà été largement étudiée. Plus récemment, des notions de profondeur pour les paramètres de *scatter* ont été définies.

Les propriétés de robustesse de la profondeur de Tukey sont déjà bien connues. Cependant, la robustesse de la profondeur scatter n'a pas encore été analysée. Dans cet exposé, nous présentons plusieurs résultats concernant la fonction de profondeur scatter et sa médiane associée, notamment la fonction d'influence et le point de rupture.

Nous poursuivons en étudiant de nouvelles propriétés, s'intéressant notamment à la médiane scatter et ses directions minimisantes. De plus, nous analysons les propriétés en échantillon fini de cette profondeur, en examinant, par exemple, la valeur maximale que cette dernière peut atteindre en fonction de la taille de l'échantillon et de la dimension.

Mots-clés. Profondeur scatter, Robustesse

Abstract. Statistical depth provides robust nonparametric tools to analyze distributions. Depth functions indeed measure the adequacy of distributional parameters to underlying probability measures. In the *location* case, the celebrated halfspace depth has been widely studied. Recently, depth notions for *scatter* parameters have been defined.

The robustness properties of the Tukey depth have already been amply discussed. However, the robustness of the scatter depth remains largely unknown. In this talk, we present several results regarding the scatter depth function and its associated scatter median, including the influence function and the breakdown point.

We continue with investigating new properties, mainly interested with the scatter median and its minimizing directions. In addition, we study finite sample properties of this depth, examining, for example, the maximum depth value depending on the sample size and dimension.

Keywords. Scatter depth, Robustness

1 Profondeur scatter

Génériquement, une fonction de profondeur $D(., P) : \theta \to D(\theta, P)$ associe à tout paramètre θ une mesure de l'adéquation de celui-ci avec la distribution de probabilité P sous-jacente. Cet outil nonparamétrique permet typiquement de construire des méthodologies robustes pour l'estimation ou le test dans des distributions paramétriques. Par exemple, dans le cas *position*, la profondeur de demi espace de Tukey (1975) est définie comme

$$HD_{\text{loc}}(\theta, P) = \inf_{||u||=1} P[u'(X - \theta) \geq 0].$$

Récemment, des notions de profondeur pour les paramètres de *dispersion* ont été définies et étudiées. Dans ce document, nous examinons principalement les propriétés de la profondeur de demi-espace de scatter introduite par Chen *et al.* (2018), dont l'expression est donnée par

$$HD_{\text{sc}}(\Sigma, P) = \inf_{||u||=1} \min \left(P[|u'(X - T_P) \leq \sqrt{u'\Sigma u}], P[|u'(X - T_P) \geq \sqrt{u'\Sigma u}] \right),$$

et où T_P est un estimateur de position. Nous considérons également la médiane de scatter associée, définie, pour $\mathcal{M}^p$ l'ensemble des matrices symétriques définies positives, comme $S_{HD}(P) = \arg\max_{\Sigma \in \mathcal{M}^p} HD_{\text{sc}}(\Sigma, P)$.

2 Robustesse

Les propriétés de robustesse de la profondeur de Tukey ont déjà largement été étudiées : fonction d'influence (Romanazzi (2001)), comportement asymptotique (Massé (2002)), *etc.* Cependant, les propriétés de robustesse de la profondeur *scatter* restent toutefois largement inconnues.

2.1 Fonction d'influence

La fonction d'influence mesure la sensibilité d'une fonctionnelle S à l'ajout d'une faible contamination sur la distribution P. Elle se définit formellement comme

$$IF(z; S(P)) = \lim_{\epsilon \to 0^+} \frac{S(P_{\epsilon,z}) - S(P)}{\epsilon} = \frac{\partial}{\partial \epsilon} S(P_{\epsilon,z}) \bigg|_{\epsilon=0},$$

où $P_{\epsilon,z} = (1-\epsilon)P + \epsilon\Delta(z)$ correspond à la distribution contaminée et $\Delta(z)$ est la distribution de Dirac en $z \in \mathbb{R}^p$. Les résultats concernant la fonction d'influence de la profondeur scatter vont dépendre de l'estimateur de position utilisé et/ou des hypothèses de continuité sur la distribution sous-jacente. Par exemple, lorsque la distribution est continue et l'estimateur de position est connu (indépendant de P), la fonction d'influence s'écrit

$$IF(z; HD_{\text{sc}}(\Sigma, P)) = I[z \in A] - HD_{\text{sc}}(\Sigma, P)$$

pour $A = A(P)$ un ensemble dépendant de la distribution P. Par contre, pour un estimateur T_P quelconque, nous trouvons, sous certaines hypothèses sur P et T_P,

$$-HD_{\text{sc}}(\Sigma, P) + g(IF(z; T_p)) \leq IF(z; HD_{\text{sc}}(\Sigma, P)) \leq 1 - HD_{\text{sc}}(\Sigma, P) + g(IF(z; T_p)).$$

Pour la médiane, il est difficile de dégager une expression générale exacte de la fonction d'influence. Nous trouvons cependant, pour des distributions elliptiques de densité radiale f et de matrice de dispersion Σ_*,

$$IF(z; S_{HD}(P)) = \frac{\Sigma_*}{2f(1)},$$

pour certaines valeurs de z. Pour des distributions plus générales, il est également possible d'obtenir des bornes sur la fonction d'influence.

2.2 Point de rupture

Le point de rupture d'une fonctionnelle est informellement défini comme le nombre d'observations à modifier avant que celle-ci n'atteigne sa ou ses bornes. Dans le cas de la profondeur, nous considérons donc, pour une jeu de données X, les points de rupture par *substitution*

$$BP(HD_{\text{sc}})(\Sigma, X)) = \min_{1 \leq m \leq n} \{\frac{m}{n} : \inf_{X^m} HD_{\text{sc}}(\Sigma, X^m) = 0\}$$

et

$$BP(S_{HD}(X)) = \min(BP^0(S_{HD}(X)), BP^\infty(S_{HD}(X))),$$

où

$$BP^0(S_{HD}(X)) = \min_{1 \leq m \leq n} \{\frac{m}{n} : \inf_{X^m} \lambda_p(S_{HD}(X^m)) = 0\}$$

et

$$BP^\infty(S_{HD}(X)) = \min_{1 \leq m \leq n} \{\frac{m}{n} : \sup_{X^m} \lambda_1(S_{HD}(X^m)) = 0\}$$

et où X^m représente X où m données ont été arbitrairement modifiées et où $\lambda_1(S)$ et $\lambda_p(S)$ désignent, respectivement, la plus grande et la plus petite valeur propre de la matrice S. Lorsque l'estimateur de position T est fixé, nous trouvons $BP(HD_{\text{sc}}(\Sigma, P)) = HD_{\text{sc}}(\Sigma, X)$ ainsi que

$$BP^0(S_{HD}(X)) \geq (\frac{M/2 - p}{n})^+ \qquad \text{et} \qquad BP^\infty(S_{HD}(X)) \geq (\frac{M/2}{n})^+.$$

Asymptotiquement, le point de rupture atteint donc, dans le pire des cas, une valeur de $\max_\Sigma HD(\Sigma, P)/2 (= 0.25$ pour des distributions elliptiques par exemple). Le cas position non fixé dépendra grandement du choix de l'estimateur.

3 Propriétés

Une seconde partie de cette présentation est dédiée à l'étude de certaines propriétés de la fonction de scatter ainsi que de sa médiane. Nous explorons par exemple les valeurs maximales et minimales que peut prendre la profondeur de la médiane, c'est à dire $HD_{\text{sc}}(S_{HD}(P), P)$, tant pour des distributions que pour un échantillon fini. Nous montrons que la borne supérieure est de 0.5 pour des distributions continues en dimension $p \geq 2$, et varie dans le cas empirique, en fonction de la taille n et de la dimension p de l'échantillon. Pour la borne inférieure, nous montrons que celle-ci est nulle en explicitant une distribution pour laquelle la médiane a une profondeur tendant vers 0.

Pour une distribution P, nous analysons aussi le comportement des directions minimisantes, à savoir l'ensemble des directions u_* telles que

$$HD_{\text{sc}}(S_{HD}(P), P) = \min\left(P[|u_*'(X - T_P) \leq \sqrt{u_*'\Sigma u_*}], P[|u_*'(X - T_P) \geq \sqrt{u_*'\Sigma u_*}]\right).$$

Ces directions doivent en effet être en nombre suffisant, qui, dans certains cas, dépend de la dimension.

Bibliographie

Tukey, J.W. (1975). Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians, pp. 523-531.

Romanazzi, M. (2001). Influence function of halfspace depth. *J. Multivariate Anal.* 77(1), pp. 138-161.

Paindaveine, D., & Van Bever, G. (2018). Halfspace depths for scatter, concentration and shape matrices. *Ann. Statist.*, 46(6B), 3276-3307.

Liu, X., Zuo, Y. & Wang, Q. (2017). Finite sample breakdown point of Tukey's halfspace median. *Sci. China Math.* 60, pp. 861-874.

Chen, M., Gao, C. & Ren, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* 46(5), pp. 1932-1960.

Prévision Conforme pour des données hiérarchiques

Guillaume Principato[1,2] & Gilles Stoltz[2,3] & Yvenn Amara-Ouali[1,2] & Yannig Goude[1,2] & Bachir Hamrouche[1] & Jean-Michel Poggi[2,4]

[1] *EDF R&D, Palaiseau, France*
[2] *Université Paris-Saclay, CNRS, Inria, Laboratoire de mathematiques d'Orsay, France*
[3] *HEC Paris, Jouy-en-Josas, France*
[4] *Universite Paris Cité, France*

Résumé. Nous considérons la prévision conforme de données multivariées, qui consiste à produire des régions de prévision basées sur les quantiles empiriques des erreurs d'estimation ponctuelle. Nous considérons en particulier le cas des données multivariées hiérarchiques, pour lesquelles certaines composantes sont des combinaisons linéaires d'autres composantes. L'intuition est que la structure hiérarchique peut être exploitée pour améliorer les régions de prévision en termes de taille pour des niveaux de couverture donnés. Nous exploitons cette intuition en incluant une étape de projection (également appelée étape de réconciliation) dans la procédure *Split Conformal Prediction* (SCP) et nous demontrons que les régions de prévision qui en résultent sont plus petites que sans l'étape de projection. Nos méthodes et leurs analyses s'appuient sur les littératures de la prévision conforme et de la réconciliation de prévisions.

Mots-clés. Prévision Conforme, Réconciliation de Prévisions, Données Hiérarchiques, Données Multivariées, Quantification de l'incertitude.

Abstract. We consider conformal prediction of multivariate data series, which consists of outputting prediction regions based on empirical quantiles of point-estimate errors. We actually consider hierarchical multivariate data series, for which some components are linear combinations of others. The intuition is that the hierarchical structure may be leveraged to improve the prediction regions in terms of their sizes for given coverage levels. We implement this intuition by including a projection step (also called reconciliation step) in the split conformal prediction (SCP) procedure and prove that the resulting prediction regions are indeed globally smaller than without the projection step. The associated strategies and their analyses rely on the literatures of both SCP and forecast reconciliation.

Keywords. Conformal Prediction, Forecast Reconciliation, Hierarchical Data, Multivariate Data, Uncertainty Quantification.

[1]mail : guillaume.principato@universite-paris-saclay.fr

1 Objectifs et Méthodologie

1.1 Données Hiérarchiques

Nous considérons le problème de régression multivarié où l'on cherche à estimer l'observation $\boldsymbol{y} \in \mathbb{R}^m$ en fonction de la variable $\boldsymbol{x} \in \mathbb{R}^d$. L'observation $\boldsymbol{y}$ est dite hiérarchique lorsque certaines de ses composantes sont obtenues par sommation des autres. Plus formellement, considerons le vecteur des n premières composantes $\boldsymbol{y}_{1:n}$ avec $2 \leq n < m$, $\boldsymbol{y}$ suit une structure hiérarchique s'il existe une matrice H de taille $m \times n$ de la forme

$$H = \begin{bmatrix} \mathrm{Id}_n \\ H_{\mathrm{sub}} \end{bmatrix} \quad \text{telle que} \quad \boldsymbol{y} = H\boldsymbol{y}_{1:n} \tag{1}$$

La matrice H encode les contraintes hiérarchiques et est appelée matrice structurelle. Deux exemples de hiérarchies avec les matrices structurelles associées sont représentés Figure 1.

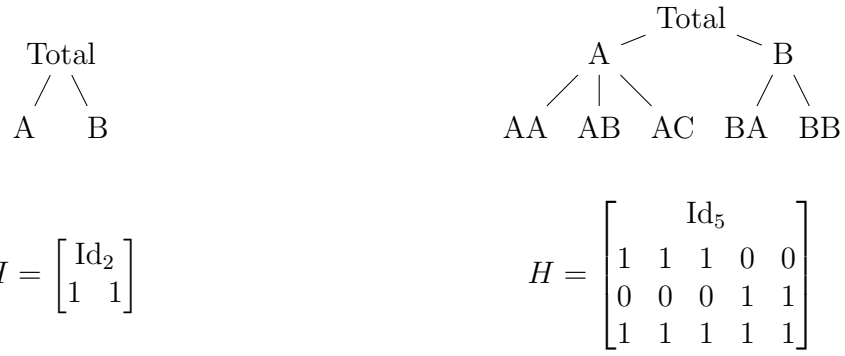


Figure 1: Exemples de structures hiérarchiques à 1 niveau (à gauche) et 2 niveaux (à droite).

Définition 1.1. Les vecteurs $\boldsymbol{u} \in \mathbb{R}^m$ qui satisfont les contraintes linéaires $\boldsymbol{u} = H\boldsymbol{u}_{1:n}$ sont dits cohérents. Le sous-espace $\mathrm{Im}(H)$ est appelé le sous-espace cohérent.

1.2 Objectifs

Nous considérons que les données i.i.d. $(\boldsymbol{x}_t, \boldsymbol{y}_t)_{1 \leq t \leq T}$ sont disponibles pour la tâche de régression (cette hypothèse sera plus tard relaxée). Nous construisons des intervalles de prévisions à partir de ce T-échantillon avec un objectif de couverture ($\star$) et un objectif de taille (ou d'efficacité) ($\star\star$).

Plus précisement, les intervalles de prévision sont de la forme $C_1 \times \ldots \times C_m$, avec $C_i : \boldsymbol{x} \in \mathbb{R}^d \mapsto C_i(\boldsymbol{x})$ une application à valeur dans les sous-ensembles de $\mathbb{R}$. Soit $(\boldsymbol{x}_{T+1}, \boldsymbol{y}_{T+1})$ une nouvelle donnée i.i.d. du T-échantillon, les intervalles de prévisions doivent, pour chaque composante $\boldsymbol{y}_{T+1,i}$, vérifier un niveau de couverture d'approximativement $1 - \alpha$:

$$\forall i \in [\![1, m]\!], \quad \mathbb{P}\big(y_{T+1,i} \in C_i(\boldsymbol{x}_{T+1})\big) \approx 1 - \alpha\,, \tag{$\star$}$$

où la probabilité $\mathbb{P}$ est par rapport à $(\boldsymbol{x}_{T+1}, \boldsymbol{y}_{T+1})$ et $(\boldsymbol{x}_t, \boldsymbol{y}_t)_{1\leq t\leq T}$. Nous désignons cette objectif comme étant la couverture composante par composante, en opposition avec la couverture jointe.

En plus de cet objectif de couverture, nous voulons des intervalles de prévision informatifs et donc souhaitons contrôler leur taille. Nous considérons dans la suite des intervalles de la forme $C_i(\boldsymbol{x}) = \big[\mu_i(\boldsymbol{x}) + a_i(\boldsymbol{x}),\, \mu_i(\boldsymbol{x}) + b_i(\boldsymbol{x})\big]$ et notons la longueur associée

$$\ell\big(C_i(\boldsymbol{x})\big) = b_i(\boldsymbol{x}) - a_i(\boldsymbol{x})\,. \tag{2}$$

Pour mesurer la taille des intervalles de prévision, nous utilisons

$$\sum_{i=1}^{m} w_i\, \ell\big(C_i(\boldsymbol{x})\big)^2 \tag{3}$$

avec $\boldsymbol{w} = (w_1, \ldots, w_m)^\top$ un vecteur de nombres positifs qui permet de pondérer les composantes selon leur importance. Nous définissons alors le second objectif :

$$\text{minimiser} \quad \mathbb{E}\left[\sum_{i=1}^{m} w_i\, \ell\big(C_i(\boldsymbol{x}_{T+1})\big)^2\right], \tag{$\star\star$}$$

avec l'espérance $\mathbb{E}$ qui est par rapport à $(\boldsymbol{x}_{T+1}, \boldsymbol{y}_{T+1})$ et $(\boldsymbol{x}_t, \boldsymbol{y}_t)_{1\leq t\leq T}$.

1.3 Prévision Conforme Hiérarchique

Nous adaptons la procédure *Split Conformal Prediction* de [3] au cadre multivarié en l'appliquant indépendamment pour chaque composante. Cette procédure consiste à décomposer le T-échantillon en deux. On note $\mathcal{D}_{\text{train}}$ et $\mathcal{D}_{\text{calib}}$ les ensembles qui contiennent leurs indices et T_{train} et T_{calib} leurs cardinalités respectives. Pour les données $(\boldsymbol{x}_t, \boldsymbol{y}_t)$ indexées par $t \in \mathcal{D}_{\text{train}}$, une fonction de régression $\widehat{\boldsymbol{\mu}} : \boldsymbol{x} \mapsto \widehat{\boldsymbol{\mu}}(\boldsymbol{x}) \in \mathbb{R}^m$ est obtenue à partir d'un algorithme de régression $\mathcal{A}$, considéré de façon boîte noire comme un paramètre de la méthode. Ensuite, les données $(\boldsymbol{x}_t, \boldsymbol{y}_t)$ indexées par $t \in \mathcal{D}_{\text{calib}}$ sont utilisées pour construire des scores de non-conformité à partir des estimations ponctuelles $\widehat{\boldsymbol{y}}_t = \widehat{\boldsymbol{\mu}}(\boldsymbol{x}_t)$. Nous choisissons les résidus (signés) $\widehat{\boldsymbol{s}}_t = \boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t$ et les traitons composante par composante en les réordonnant avec l'extension de la notion de statistiques d'ordre au cas composante par composante :

$$\forall i \in [\![1, m]\!], \qquad \widehat{s}_{(1),i} \leq \ldots \leq \widehat{s}_{(T_{\text{calib}}),i}$$

et définissons également $\widehat{s}_{(0),i} = -\infty$ and $\widehat{s}_{(T_{\text{calib}}+1),i} = +\infty$. L'intervalle de prévision résultant pour $\widehat{\boldsymbol{y}}_{T+1}$ basé sur la variable $\boldsymbol{x}_{T+1}$ est alors

$$\widehat{C}_i(\boldsymbol{x}_{T+1}) = \left[\widehat{\mu}_i(\boldsymbol{x}_{T+1}) + \widehat{s}_{\left(\lfloor (T_{\text{calib}}+1)\alpha/2 \rfloor\right),i}\,,\; \widehat{\mu}_i(\boldsymbol{x}_{T+1}) + \widehat{s}_{\left(\lceil (T_{\text{calib}}+1)(1-\alpha/2) \rceil\right),i}\right] \tag{4}$$

avec $1-\alpha \in [0,1]$ est le niveau de couverture voulu et $\widehat{\mu}_i(\boldsymbol{x}_{T+1})$ est la i–ième composante de l'estimation ponctuelle $\widehat{\boldsymbol{\mu}}(\boldsymbol{x}_{T+1})$. Nous synthétisons cette procédure dans l'Algorithme 1.

Algorithme 1 Benchmark SCP multivariée avec des scores signés

Paramètres : niveau de couverture $1-\alpha$; algorithme de régression $\mathcal{A}$; partition de $[T]$ en $\mathcal{D}_{\text{train}}$ et $\mathcal{D}_{\text{calib}}$ de cardinalités respectives T_{train} et T_{calib}
Construire le régresseur $\widehat{\boldsymbol{\mu}}(\,\cdot\,) = \mathcal{A}\big((\boldsymbol{x}_t, \boldsymbol{y}_t)_{t\in\mathcal{D}_{\text{train}}}\big)$
Définir $\widehat{\boldsymbol{\mu}}(\,\cdot\,) = \big(\widehat{\mu}_1(\,\cdot\,), \ldots, \widehat{\mu}_m(\,\cdot\,)\big)$
Pour $t \in \mathcal{D}_{\text{calib}}$ **soient** $\widehat{\boldsymbol{y}}_t = \widehat{\boldsymbol{\mu}}(\boldsymbol{x}_t)$ and $\widehat{\boldsymbol{s}}_t = \boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t$
Pour chaque $i \in [\![1, m]\!]$ **faire**
    Ordonner $(\widehat{s}_{t,i})_{t\in\mathcal{D}_{\text{calib}}}$ en $\widehat{s}_{(1),i} \leq \ldots \leq \widehat{s}_{(T_{\text{calib}}),i}$ et définir $\widehat{s}_{(0),i} = -\infty$ et $\widehat{s}_{(T_{\text{calib}}+1),i} = +\infty$
    Soient $\widehat{q}^{(i)}_{\alpha/2} = \widehat{s}_{\left(\lfloor (T_{\text{calib}}+1)\alpha/2 \rfloor\right),i}$ et $\widehat{q}^{(i)}_{1-\alpha/2} = \widehat{s}_{\left(\lceil (T_{\text{calib}}+1)(1-\alpha/2) \rceil\right),i}$
    Définir $\widehat{C}_i(\,\cdot\,) = \Big[\widehat{\mu}_i(\,\cdot\,) + \widehat{q}^{(i)}_{\alpha/2},\ \widehat{\mu}_i(\,\cdot\,) + \widehat{q}^{(i)}_{1-\alpha/2}\Big]$
Retourner $\widehat{C}_1(\boldsymbol{x}_{T+1}), \ldots, \widehat{C}_m(\boldsymbol{x}_{T+1})$

Cette procédure est considérée comme *benchmark* car elle n'exploite pas la structure hiérarchique présente dans les données. Nous proposons alors d'inclure une étape de projection sur $\text{Im}(H)$ (également appelée étape de réconciliation) dans la procédure ce qui défini alors l'Algorithme 2. Le concept de réconciliation est apparu dans [1] pour améliorer les prévisions ponctuelles en exploitant la structure hiérarchique tandis que la réconciliation sous l'angle des projections sur $\text{Im}(H)$ a été étudiée par [4].

Algorithme 2 Hierarchique SCP multivariée avec des scores signés

Paramètres : niveau de couverture $1-\alpha$; algorithme de régression $\mathcal{A}$; partition de $[T]$ en $\mathcal{D}_{\text{train}}$ et $\mathcal{D}_{\text{calib}}$ de cardinalités respectives T_{train} et T_{calib}; matrice P
Construire le régresseur $\widehat{\boldsymbol{\mu}}(\,\cdot\,) = \mathcal{A}\big((\boldsymbol{x}_t, \boldsymbol{y}_t)_{t\in\mathcal{D}_{\text{train}}}\big)$
Définir $P\widehat{\boldsymbol{\mu}}(\,\cdot\,) = \widetilde{\boldsymbol{\mu}}(\,\cdot\,) = \big(\widetilde{\mu}_1(\,\cdot\,), \ldots, \widetilde{\mu}_m(\,\cdot\,)\big)$
Pour $t \in \mathcal{D}_{\text{calib}}$ **soient** $\widetilde{\boldsymbol{y}}_t = \widetilde{\boldsymbol{\mu}}(\boldsymbol{x}_t)$ and $\widetilde{\boldsymbol{s}}_t = \boldsymbol{y}_t - \widetilde{\boldsymbol{y}}_t$
Pour chaque $i \in [\![1, m]\!]$ **faire**
    Ordonner $(\widetilde{s}_{t,i})_{t\in\mathcal{D}_{\text{calib}}}$ en $\widetilde{s}_{(1),i} \leq \ldots \leq \widetilde{s}_{(T_{\text{calib}}),i}$ et définir $\widetilde{s}_{(0),i} = -\infty$ et $\widetilde{s}_{(T_{\text{calib}}+1),i} = +\infty$
    Soient $\widetilde{q}^{(i)}_{\alpha/2} = \widetilde{s}_{\left(\lfloor (T_{\text{calib}}+1)\alpha/2 \rfloor\right),i}$ et $\widetilde{q}^{(i)}_{1-\alpha/2} = \widetilde{s}_{\left(\lceil (T_{\text{calib}}+1)(1-\alpha/2) \rceil\right),i}$
    Définir $\widetilde{C}_i(\,\cdot\,) = \Big[\widetilde{\mu}_i(\,\cdot\,) + \widetilde{q}^{(i)}_{\alpha/2},\ \widetilde{\mu}_i(\,\cdot\,) + \widetilde{q}^{(i)}_{1-\alpha/2}\Big]$
Retourner $\widetilde{C}_1(\boldsymbol{x}_{T+1}), \ldots, \widetilde{C}_m(\boldsymbol{x}_{T+1})$

Pour faciliter la distinction des deux méthodes, nous noterons avec $\widehat{\cdot}$ les quantités correspondant à la méthode *benchmark* (Algorithme 1) et avec $\widetilde{\cdot}$ les quantités correspondant à la méthode hiérarchique (Algorithme 2).

2 Résultats théoriques

En prévision conforme, une attention particulière est portée au fait d'obtenir des résultats de couverture avec des hypothèses minimales. Les hypothèses qui vont servir à établir les résultats de cette section s'inscrivent dans cette ligne. En particulier, aucune hypothèse n'est faite sur l'algorithme de régression $\mathcal{A}$ qui peut donc être n'importe quel modèle de prévision boîte noire.

Hypothèse 2.1. Les scores de non-conformité $\widehat{\boldsymbol{s}}_t = \boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t$, pour $t \in \mathcal{D}_{\text{calib}} \cup \{T+1\}$ de l'Algorithme 1 sont i.i.d. (ou échangeables). C'est en particulier le cas lorsque $(\boldsymbol{x}_t, \boldsymbol{y}_t)_{1 \leq t \leq T+1}$ sont i.i.d (ou échangeables).

Théorème 2.2 (Couverture)**.** *Fixons $\alpha \in [0,1]$. L'Algorithme 2, utilisé avec n'importe quel Algorithme de régression $\mathcal{A}$ et n'importe quelle matrice P telle que $PH = H$, assure que sous l'Hypothèse 2.1 (scores i.i.d.), on a :*

$$\forall i \in [\![1, m]\!], \quad \mathbb{P}\big(y_{T+1,i} \in \widetilde{C}_i(\boldsymbol{x}_{T+1})\big) \geq 1 - \alpha\,.$$

De plus, si les scores $\big(\widehat{\boldsymbol{s}}_t\big)_{t \in \mathcal{D}_{calib} \cup \{T+1\}}$ sont presque sûrement distincts, alors

$$\forall i \in [\![1, m]\!], \quad \mathbb{P}\big(y_{T+1,i} \in \widetilde{C}_i(\boldsymbol{x}_{T+1})\big) \leq 1 - \alpha + \frac{2}{T_{calib} + 1}\,.$$

La condition $PH = H$ pour la matrice P dans l'Algorithme 2 permet d'inclure l'Algorithme 1 dans le Théorème 2.2. En effet, c'est un cas particulier en choisissant $P = \mathrm{Id}_m$. Les résultats qui vont suivre requièrent des hypothèses plus restrictives que les hypothèses habituelles en prévision conforme. Elles restent cependant assez peu restrictives et sont classiques en réconciliation.

Hypothèse 2.3. Les scores de non-conformité (i.i.d.) $\widehat{\boldsymbol{s}}_t = \boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t$, pour $t \in \mathcal{D}_{\text{calib}}$, de l'Algorithme 1 suivent une distribution elliptique [2] dont les paramètres sont inconnus.

Théorème 2.4 (Longueur – version faible)**.** *Soit $\boldsymbol{w} \in \mathbb{R}^m$ un vecteur de nombres positifs. Sous les Hypothèses 2.1 et 2.3 (scores i.i.d. suivant une distribution elliptique), la version hiérarchique de SCP (Algorithme 2) exécutée avec $P = P_{\boldsymbol{w}}$, où*

$$P_{\boldsymbol{w}} := H\big(H^\top \operatorname{diag}(\boldsymbol{w}) H\big)^{-1} H^\top \operatorname{diag}(\boldsymbol{w})\,, \tag{5}$$

produit des intervalles de prévision plus efficaces que ceux de l'Algorithme 1 dans le sens où

$$\mathbb{E}\left[\sum_{i=1}^{m} w_i\ \ell\big(\widetilde{C}_i(\boldsymbol{x}_{T+1})\big)^2\right] \leq \mathbb{E}\left[\sum_{i=1}^{m} w_i\ \ell\big(\widehat{C}_i(\boldsymbol{x}_{T+1})\big)^2\right].$$

Hypothèse 2.5. Les scores de non-conformité (i.i.d.) $\widehat{\boldsymbol{s}}_t = \boldsymbol{y}_t - \widehat{\boldsymbol{y}}_t$, pour $t \in \mathcal{D}_{\text{calib}}$, de l'Algorithme 1 ont un moment du second ordre borné. On note alors Σ leur matrice (définie positive) de variance.

Théorème 2.6 (Longueur – version forte)**.** *Sous les Hypothèses 2.1, 2.3, et 2.5 (scores i.i.d. suivant une distribution elliptique qui admet un moment du second ordre borné), la version hiérarchique de SCP (Algorithme 2) exécutée avec $P = P_{\Sigma^{-1}}$, où*

$$P_{\Sigma^{-1}} := H\big(H^\top \Sigma^{-1} H\big)^{-1} H^\top \Sigma^{-1}\,, \tag{6}$$

produit des intervalles de prévision plus efficaces qu'avec n'importe quel autre choix de matrice de projection sur $\mathrm{Im}(H)$ *:*

$$\forall i \in [\![1, m]\!], \quad \mathbb{E}\Big[\ell\big(\widetilde{C}_i^\star(\boldsymbol{x}_{T+1})\big)^2\Big] \leq \mathbb{E}\Big[\ell\big(\widetilde{C}_i(\boldsymbol{x}_{T+1})\big)^2\Big]\,.$$

Corollaire 2.7. *Dans le même cadre et avec les mêmes hypothèses que pour le Théorème 2.6, nous avons également*

$$\forall i \in [\![1, m]\!], \quad \mathbb{E}\Big[\ell\big(\widetilde{C}_i^\star(\boldsymbol{x}_{T+1})\big)^2\Big] \leq \mathbb{E}\Big[\ell\big(\widehat{C}_i(\boldsymbol{x}_{T+1})\big)^2\Big]\,,$$

où $\widehat{C}_1(\boldsymbol{x}_{T+1}) \times \ldots \times \widehat{C}_m(\boldsymbol{x}_{T+1})$ *sont les intervalles de prévision obtenus par l'Algorithme 1.*

3 Conlusion

Nous venons de présenter des résultats théoriques en faveur de l'utilisation de Hiérarchique SCP avec des scores signés pour quantifier l'incertitude autour de la prévision de données hiérarchiques. Pour plus de détails, notamment sur l'implémentation de la méthode en pratique et ses performances sur des données réelles, nous renvoyons le lecteur vers [5].

References

[1] Rob J. Hyndman, Roman A. Ahmed, George Athanasopoulos, and Han Lin Shang. Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9):2579–2589, 2011.

[2] Tonu Kollo and Dietrich von Rosen. *Advanced Multivariate Statistics with Matrices*. Mathematics and Its Applications. Springer, 2005.

[3] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[4] Anastasios Panagiotelis, George Athanasopoulos, Puwasala Gamakumara, and Rob J. Hyndman. Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting*, 37(1):343–359, 2021.

[5] Guillaume Principato, Gilles Stoltz, Yvenn Amara-Ouali, Yannig Goude, Bachir Hamrouche, and Jean-Michel Poggi. Conformal prediction for hierarchical data. *arXiv preprint arXiv:2411.13479*, 2025.

Statistique non paramétrique

Estimateur récursif de la densité par noyau asymétrique sous dépendance η-faible

Smail Adjabi [1], Mohamed Khemici [2] & Nabil Zougab [3]

[1] *Unité de Recherche LaMOS, Université de Bejaia, Algérie, smail.adjabi@univ-bejaia.dz*
[2] *Unité de Recherche LaMOS, Université de Bejaia, Algérie, mohamed.khemici@univ-bejaia.dz*
[3] *Unité de Recherche LaMOS, Université Bejaia, Algérie, nabil.zougab@univ-bejaia.dz*

Résumé. On présente l'estimateur récursif de la fonction densité par noyau asymétrique (Gamma, Inverse Gamma et Réciproque Inverse Gamma) sous dépendance η-faible. L'erreur quadratique moyenne intégrée (MISE) est établie comme une propriété globale. Une étude de simulation examine les performances de l'estimateur et compare ses performances avec l'estimateur non récursif pour une dépendance η-faible.
Mots-clés. Estimateur récursif, Dépendance η-faible, Noyau, Paramètre de lissage, Erreur quadratique moyenne intégrée.

Abstract. We presente the recursive asymetric kernel(Gamma, Inverse Gamma and Reciprocal Inverse Gamma) density estimation under η-weak dependence. The mean integrated squared error (MISE) is established as a global property. A simulation study investigates the performance of some asymmetric kernel recursive kernel density estimators and compares their performance with the non-recursive density estimators for η-weak dependence.

Keywords. Recusive estimator η-weak dependence, kernel, Smoothing parameter, Mean integrated squared error.

1 Introduction

Dedecker et al. (2007), Mokkadem et al (2009), et Amiri (2010) ont proposé un estimateur récursif de la fonction densité par noyau symétrique sous une dépendance η-faible. On élargi cet estimateur aux noyaux asymétriques, en considérant exclusivement le noyau Gamma(G) de Chen (2000), le noyau Inverse Gaussien (IG) et le noyau Réciproque Inverse Gaussien (RIG) de Scaillet (2004). Nous obtenons les propriétés statistiques de l'estimateur (biais, variance et l'erreur quadratique moyenne intégrée (MISE)) pour les données dépendantes d'un processus η-faible stationnaire en utilisant les noyaux G, IG et RIG. Une étude de simulation de Monte Carlo basée sur des données dépendantes d'un modèle autorégressif de durée conditionnelle (ACD) et d'un modèle de volatilité stochastique (SV) est réalisée en utilisant les noyaux asymétriques G, IG et RIG. Les performances de l'estimateur récursif à noyau asymétrique sont comparées avec ceux de l'estimateur non récursif sous une dépendance η faible sur des échantillons finis au sens du critère de l'erreur quadratique intégrée (ISE).

2 Estimateur récursif de la densité pour une dépendance η-faible

2.1 Estimateur

Soit $X_1, X_2, \cdots, X_n$ une suite stationnaire de variables aléatoires réelles faiblement dépendantes de fonction densité f inconnue. L'estimateur récursif à noyau asymétrique pour une dépendance η-faible de $f(x)$ est de la forme:

$$\hat{f}_n(x) = \frac{1}{\sum_{i=1}^{n} h_i} \sum_{i=1}^{n} h_i K_{x,h_i}(X_i), \tag{1}$$

où h_i est le paramètre de lissage et K_{x,h_i} est le noyau asymétrique continu récursif dépendant de la cible x et du paramètre de lissage h_i.
Wu (2005) a proposé des mesures de dépendance qui fournissent un cadre naturel pour une théorie des limites pour les processus stationnaires. La suite $(X_t)_{t\in N}$ est η-faiblement dépendante (voir Dedecker et al (2007)), s'il existe une séquence décroissante $(\eta_k)_{k\in N}$ tendant vers 0 à l'infini vérifiant les deux conditions:
Condition 1. Dedecker et al (2007)

$$|\mathrm{Cov}(g_1(X_{s_1}, ..., X_{s_u}), g_2(X_{t_1}, ..., X_{t_u}))| \leq (u \parallel g_2 \parallel_\infty Lip(g_1) + v \parallel g_1 \parallel_\infty Lip(g_2))\eta_k,$$

avec

- $s_1 \leq ... \leq s_u \leq s_u + k \leq t_1 \leq ... \leq t_v$ où $u, v \in N^*, (s_1, ..., s_u) \in N^u$ and $(t_1, ..., t_v) \in N^*$
- $g_1 : R^{+^u} \to R, g_2 : R^{+^v} \to R$, sont deux fonctions bornées.

Condition 2. Dedecker et al (2007)
$Lip(g_1) < \infty$, où $Lip(g_1) = \sup_{x\neq y} \frac{|g(x)-g(y)|}{\|x-y\|_l}$, $l = 1, 2$ et $\parallel X \parallel_l = \sum_{i=1}^{u} |X_i|$.

Afin d'établir les propriétés de l'estimateur, nous supposerons les hypothèses suivantes:
(H.1): $B_{n,r} := \frac{1}{n}\sum_{i=1}^{n}(\frac{h_i}{h_n})^r \to B_r < \infty, n \to \infty$
(H.2): $h_n \downarrow 0$ et $h_i = h_n \to 0$ quand $n \to \infty$
(H.3): f est deux fois différentiable.

2.2 Biais de l'estimateur

Considérons l'estimateur récursif de la fonction densité défini par (1). En supposant que les hypothèses (H.1), (H.2) et (H.3) sont vérifiées, le biais de l'estimateur récursif pour chaque noyau asymétrique G, IG et RIG est de la forme:

$$Biais(\hat{f}_G^n(x)) = \frac{h_n B_2}{B_1}\left[f'(x) + \frac{x}{2}f"(x)\right];$$

$$Biais(\hat{f}_{IG}^{n}(x)) = \frac{h_n B_2}{B_1}\left[\frac{1}{2}x^3 f"(x)\right];$$
$$Biais(\hat{f}_{RIG}^{n}(x)) = \frac{h_n B_2}{B_1}\left[\frac{1}{2}x f"(x)\right].$$

2.3 Variance de l'estimateur

Considérons l'estimateur récursif de la fonction densité défini par (1). En supposant que les hypothèses (H.1), (H.2) et (H.3) sont vérifiées, la variance de l'estimateur récursif est de la forme:

$$Var(\hat{f}_n(x)) = \frac{1}{(\sum_{i=1}^{n} h_i)^2}\sum_{i=1}^{n} h_i^2 Var(K_{x,h_i}(X_i)) + 2\left(\sum_{i=1}^{n} h_i\right)^{-2}\sum_{i>j} A_{i,j},$$

$$\text{où } A_{i,j} = COV(h_i K_{x,h_i}(X_i), h_j K_{x,h_j}(X_j)).$$

La variance de l'estimateur récursif pour chaque noyau asymétrique G, IG et RIG est alors:

$$Var(\hat{f}_G^n(x)) = \frac{B_{\frac{3}{2}}}{nh_n^{\frac{1}{2}}B_1^2}\left[\frac{1}{2\sqrt{\pi}}x^{\frac{-1}{2}}f(x)\right];$$

$$Var(\hat{f}_{IG}^{n}(x)) = \frac{B_{\frac{3}{2}}}{nh_n^{\frac{1}{2}}B_1^2}\left[\frac{1}{2\sqrt{\pi}}x^{\frac{-3}{2}}f(x)\right];$$

$$Var(\hat{f}_{RIG}^{n}(x)) = \frac{B_{\frac{3}{2}}}{nh_n^{\frac{1}{2}}B_1^2}\left[\frac{1}{2\sqrt{\pi}}x^{\frac{-1}{2}}f(x)\right].$$

2.4 Erreur Quadratique Moyenne intégrée (MISE) de l'estimateur

Considérons l'estimateur récursif de la fonction densité défini par (1). En supposant que les hypothèses (H.1), (H.2) et (H.3) sont vérifiées, l'erreur quadratique moyenne intégrée $MISE(\hat{f}_n) = \int(Biais^2(\hat{f}_n(x)) + Var(\hat{f}_n(x)))dx$ de l'estimateur récursif pour chaque noyau asymétrique G, IG et RIG est de la forme:

$$MISE(\hat{f}_G^n) = \frac{h_n^2 B_2^2}{B_1^2}\int_0^\infty\left[f'(x) + \frac{x}{2}f"(x)\right]^2 dx + \frac{B_{\frac{3}{2}}}{nh_n^{\frac{1}{2}}B_1^2}\left[\frac{1}{2\sqrt{\pi}}\int_0^\infty x^{\frac{-1}{2}}f(x)\right]dx;$$

$$MISE(\hat{f}_{IG}^{n}) = \frac{h_n^2 B_2^2}{B_1^2}\int_0^\infty\left[\frac{x^3}{2}f"(x)\right]^2 dx + \frac{B_{\frac{3}{2}}}{nh_n^{\frac{1}{2}}B_1^2}\left[\frac{1}{2\sqrt{\pi}}\int_0^\infty x^{\frac{-3}{2}}f(x)\right]dx;$$

$$MISE(\hat{f}_{RIG}^{n}) = \frac{h_n^2 B_2^2}{B_1^2}\int_0^\infty\left[\frac{x}{2}f"(x)\right]^2 dx + \frac{B_{\frac{3}{2}}}{nh_n^{\frac{1}{2}}B_1^2}\left[\frac{1}{2\sqrt{\pi}}\int_0^\infty x^{\frac{-1}{2}}f(x)\right]dx.$$

3 Simulation

L'objectif est d'évaluer les performances de l'estimateur récursif proposé en utilisant les noyaux asymétriques G, IG, RIG. Pour cela, une étude de simulation est réalisée, basée sur les échantillons de variables aléatoires dépendantes générées à partir d'un modèle de durée conditionnelle autorégressive (DCA) et d'un modèle de volatilité stochastique (VS) (Bouezmarni et Rombouts (2010)). Nous avons comparé l'estimateur récursif avec l'estimateur non récursif pour les noyaux asymétriques G, IG, RIG. Nous effectuons 100 réplications pour une dépendance faible η et pour des tailles d'échantillon de $50, 100, 300$ et 500. Le paramètre de lissage $h_n = hn^{-2/5}$, où h est estimé par la méthode de validation croisée non biaisé (UCV) Khemici et al (2024), qui consiste à minimiser un estimateur sans biais de $ISE(h) = \int \left(\widehat{f}(x) - f(x)\right)^2 dx$. Cette évaluation des performances est basée sur le critère de comparaison Integrate Squared Error (ISE) défini par : $ISE = \int \left(\widehat{f}(x) - f_{emp}(x)\right)^2 dx$ où $\widehat{f}$ est l'estimateur récursif ou non récursif du noyau et f_{emp} est la fonction de densité empirique.

Modèle ACD	Modèle SV
	$X_i = exp(y_i)$
$x_i = \Psi_i \xi_i$	$y_i = exp(x_i/2)\mu_i$
$\Psi_i = w + \alpha_1 x_{i-1} + \alpha_2 \Psi_{i-1}$	$x_i = w + \alpha_1 x_{i-1} + \alpha_2 \varepsilon_i$

Table 1: Modèles ACD et SV

Pour le modèle ACD, on considère deux modèles pour ξ_i

- Modèle A: $\xi_i \sim Gamma(4.92, 4.92)$;
- Modèle B: $\xi_i \sim Exp(1.02)$.
  Les paramètres de faible dépendance sont: $w = 0.153$, $\alpha_1 = 0.11$ et $\alpha_2 = 0.80$.

Pour le Modèle SV on considère le modèle Gaussien pour ϵ_i et μ_i comme

- Modèle C: $\epsilon_i \sim \mathcal{N}(0.1, 0.4)$ et $\mu_i \sim \mathcal{N}(0.1, 0.4)$.
  Les paramètres de faible dépendance sont: $w = 0.01$, $\alpha_1 = 0.95$ et $\alpha_2 = 0.30$.

A partir du tableau 2, on remarque que:

- Les valeurs du ISE diminuent à mesure que la taille de l'échantillon augmente et ceci pour les trois modèles et quel que soit le noyau utilisé.
- Dans la majorité des modèles, l'estimateur récursif est meilleur que l'estimateur non récursif au sens du critère ISE.

Modèle	Noyau	Méthode	n=50	n=100	n=300	n=500
A	G	Non recursive	0.063427790	0.03539776	0.021836248	0.0179675524
		Recursive	0.032876571	0.01336442	0.008596550	0.0070844031
	IG	Non recursive	0.020669161	0.01054410	0.009251655	0.0086498543
		Recursive	0.019957046	0.01285972	0.008967747	0.00847619
	RIG	Non recursive	0.129105060	0.06509415	0.050947447	0.0376416706
		Recursive	0.048647543	0.02546460	0.016718525	0.00156540636
B	G	Non recursive	0.507411709	0.32273986	0.298105766	0.2535550185
		Recursive	0.008883898	0.00389148	0.002826125	0.0020592013
	IG	Non recursive	0.213777008	0.19885718	0.178260378	0.1438567459
		Recursive	0.168548507	0.08320118	0.047534294	0.0460701694
	RIG	Non recursive	0.115275146	0.11515548	0.114981887	0.1070559947
		Recursive	0.571394253	0.11536935	0.114248428	0.01127560569
C	G	Non recursive	0.165448265	0.08265119	0.052594029	0.0480639442
		Recursive	0.178527985	0.10131835	0.059226036	0.0510278658
	IG	Non recursive	0.004185259	0.0 0213982	0.002146376	0.0019387301
		Recursive	0.002448519	0.00119247	0.000944777	0.0007139989
	RIG	Non recursive	0.045217854	0.02048072	0.015438055	0.0113052255
		Recursive	0.048231065	0.02259201	0.014091676	0.0115732563

Table 2: Comparaison des ISE des estimateurs à noyau (G, IG et RIG) récursif et non récursif pour modèles ACD et SV.

4 Conclusion

Nous avons étudié l' estimateur à noyau asymétrique (G, IG et RIG) récursif pour des données positives η-faiblement dépendantes stationnaires. Les propriétés statistiques(biais, variance et l'erreur quadratique moyenne intégrée (MISE) comme propriété globale) sont établies. Les performances de l'estimateur de la fonction densité à noyau récursif et non récursif sont comparés à travers une étude de simulation des modèles ACD et SV. L'estimateur récursif à noyau asymétrique est plus performant que l'estimateur non récursif au sens de l'erreur quadratique intégrée (ISE).

Bibliographie

Amiri, A. (2010). *Rates of almost sure convergence of families of recursive kernel estimators*, Annales de l'ISUP, 54, 3, pp. 3-24.

Bouezmarni, T. and Rombouts, J. (2010). *Nonparametric density estimation for positive time series*, Computational Statistics & Data Analysis, 54,2, pp. 245-261.

Chen, S. (2000). *Probability density function estimation using gamma kernels*, Annals of the Institute of Statistical Mathematics, 52, pp. 471- 480.

Dedecker, J. Doukhan, P. Lang, G. José Raphael, L. R. Louhichi, S. and Prieur, C. (2007).

Weak dependence. Springer.

Mezhoud, K. Mohdeb, Z. and Louhichi, S. (2014).*Recursive kernel estimation of the density under η-weak dependence*. Journal of the Korean Statistical Society, 43, 3, pp. 403-414.

Mokkadem, A. Pelletier, M. and Slaoui, Y. (2009). *The stochastic approximation method for the estimation of a multivariate probability density*. Journal of Statistical Planning and Inference, 139 ,7 , pp. 2459-2478.

Scaillet, O. (2004). *Density estimation using inverse and reciprocal inverse Gaussian kernels*, Nonparametric statistics, 16,1-2, pp. 217-226.

Wu, W.B. (2005). *Nonlinear system theory: Another look at dependence*. Proceedings of the National Academy of Sciences,102,40, pp. 14150-14154.

Heavy-tailed and Horseshoe priors for regression and sparse Besov rates

Sergios Agapiou [1] & Ismaël Castillo [2]& Paul Egels [2]

[1] *Department of Mathematics and Statistics, University of Cyprus, Nicosia, Cyprus. agapiou.sergios@ucy.ac.cy*

[2] *Sorbonne Université, LPSM; Paris, France. ismael.castillo@sorbonne-universite.fr, paul.egels@sorbonne-universite.fr.*

Résumé. La grande diversité de fonctions rencontrées dans les problématiques de traitement du signal et d'inférence non paramétrique nécessite la conception de méthodes suffisamment flexibles pour garantir des performances optimales ou quasi-optimales sur une large gamme de classes fonctionnelles. Par exemple, en considérant une notion générale de régularité, possiblement inhomogène comme c'est le cas dans des classes de Besov accompagnées d'un panel de différentes fonctions de perte. Dans ce travail, nous montrons que l'utilisation d'un a priori indépendant et à queues lourdes sur les coefficients d'une base orthonormée de L^2 donnée, introduit dans [2] sous le nom d'Oversmoothed Heavy-Tailed (OT) prior, assure que la loi a posteriori associée possède la flexibilité et l'optimalité requises. Nous étudions également, pour la première fois dans un contexte non paramétrique, les a priori de type horseshoe. De plus, nous fournissons une implémentation de la méthode proposée et illustrons nos résultats au travers de simulations.

Mots-clés. Bayesien non-paramétrique, analyse fréquentiste des méthodes Bayesiennes, prior à queues lourdes, loi horseshoe, espaces de Besov, vitesse Besov-sparse

Abstract. The large variety of possible functions encountered in statistics for signal processing calls for methods that are flexible enough to achieve optimal or near-optimal performance over a wide variety of functional classes, such as Besov balls, as well as over a large array of loss functions. In this work, we show that a class of heavy-tailed prior distributions on basis function coefficients introduced in [2] called Oversmoothed heavy-Tailed (OT) priors satisfies this requirements. We investigate also the case of horseshoe distributions, for the first time in the context of nonparametrics, and show that they fit into this framework. We also provide implementation of the proposed method and illustrate our results through a simulations study.

Keywords. Bayesian nonparametrics, frequentist analysis of posterior distributions, heavy-tailed prior distributions, horseshoe prior, Besov spaces, sparse Besov rates.

1 Introduction

Gaussian white noise model. For $n \geq 1$ and a given true regression function $f_0 \in L^2 := L^2([0,1])$ with coefficients $f_{0,k} = \langle f_0, \varphi_k \rangle$ into a given orthonormal basis $\{\varphi_k : k \geq 1\}$,

consider data generated from the (projected) GWN model

$$X_k | f_{0,k} \sim \mathcal{N}(f_{0,k}, 1/n), \tag{1}$$

independently in $k \geq 1$. In the following we will write $X = X^{(n)} = (X_1, X_2, \dots)$ the corresponding sequence of observations.

Frequentist analysis of posterior distributions. Consider data X generated from the sequence model (1) with true regression function $f_0 = (f_{0,k}) \in \mathcal{F}$, where $\mathcal{F}$ is the parameter space. In the following, this will be written as $X = X^{(n)} \sim P_{f_0}^{(n)}$. The reconstruction of f_0 from the data X will be conducted through a Bayesian analysis : from a prior distribution Π on $\mathcal{F}$ to be chosen below, one constructs a data-dependent probability measure on $\mathcal{F}$ called posterior distribution and given, for any measurable $B \subset \mathcal{F}$, by

$$\Pi[B \,|\, X] := \frac{\int_B \exp\{-\frac{n}{2}\sum_{k\geq 1}(X_k - f_k)^2\}\, d\Pi(f)}{\int \exp\{-\frac{n}{2}\sum_{k\geq 1}(X_k - f_k)^2\}\, d\Pi(f)}. \tag{2}$$

Heavy tailed series priors. For a function $f \in L^2$, we define a prior Π on f by drawing the coefficients f_k independently using heavy-tailed density functions. We are interested in heavy-tailed priors as in [2] that take on the form

$$f_k = \sigma_k \zeta_k, \tag{3}$$

where $\{\sigma_k\}$ are deterministic decay coefficients and ζ_k are i.i.d. random variables with a given heavy-tailed density h on $\mathbb{R}$, we assume that h satisfies the following conditions: there exist constants $c_1, c_2 > 0$ and $\kappa \geq 0$ such that

(H1) h is symmetric, positive, and decreasing on $[0, +\infty)$,

(H2) for all $x \geq 0$,
$$\log\left(1/h(x)\right) \leq c_1(1 + \log^{1+\kappa}(1 + x)),$$

(H3) for all $x \geq 1$,
$$\overline{H}(x) := \int_x^{+\infty} h(u)du \leq \frac{c_2}{x}.$$

The choice of decay is made to be asymptotically smaller than any polynomial decay, called *Oversmoothed Heavy Tailed* or OT priors,

$$\text{for all } k \geq 1, \qquad \sigma_k = e^{-(\log k)^{1+\nu}}, \tag{4}$$

where $\nu > 0$. The idea behind the prior (3) is to shrink the values of the coefficients by a deterministic small scaling factor σ_k but also ensure that the prior will catch the high-valued coefficients with non-trivial probability thanks to the heavy-tails of ζ_k.

Horseshoe priors. We give a specific example for a prior of the form (3) that is commonly used in practice, the Horseshoe prior [3] with parameter $\sigma_k > 0$,

$$f_k \sim HS(\sigma_k) \qquad \text{independently across } k \geq 1. \tag{5}$$

For any $\tau > 0$, the distribution of $f \sim HS(\tau)$ is given by $\lambda \sim C^+(0,1)$ and $f|\lambda \sim \mathcal{N}(0, \tau\lambda^2)$ where $C^+(0,1)$ is the half-Cauchy distribution with location parameter 0 and scale 1. The Horseshoe distribution has a density featuring a pole at zero and Cauchy-like tails. A variation of the Horseshoe prior that is of common use in high-dimensional models is a truncated version with uniform rescaling (with respect to k), i.e. for some $\tau > 0$,

$$f_k \overset{i.i.d}{\sim} HS(\tau) \text{ for } k \leq n \quad \text{and} \quad f_k = 0 \text{ for } k > n. \tag{6}$$

This prior matches the definition of the heavy-tailed prior (3) up until the truncation point $k = n$ after which it assigns all coefficients to zero. The uniform rescaling τ is typically of the order of a negative power of n, see Theorem 1 below.

Smoothness of the truth. To study the rate at which the posteriors defined by (2) concentrates around the truth f_0, we need to specify a parameter space. In Sections 2 the concentration towards f_0 will be given in L^2 loss, a suited parameter space is the Hilbert-Sobolev-type ball. Recalling that $f_k = \langle f, \varphi_k \rangle$, for $\beta, F > 0$, denote

$$\mathcal{S}^\beta(F) = \Big\{ f \in L^2([0,1]) \,:\, \sum_{k \geq 1} k^{2\beta} |f_k|^2 \leq F^2 \Big\}, \tag{7}$$

for a suitable choice of orthonormal basis $\{\varphi_k\}$, $\mathcal{S}^\beta(F)$ corresponds to balls of L^2 functions with β derivatives in L^2. A more general study of contraction of the posterior under other L^p losses over Besov spaces is conducted in Section 3

2 Contraction rates in L^2-loss

2.1 Upper bound for the OT prior

Theorem 1. *Consider $f_0 \in \mathcal{S}^\beta(F)$ and Π the OT series prior with independent prior on coefficients (f_k) given by* (3) *(in particular, the prior* (5) *is admissible) and decay sequence given by* (4)*. As $n \to \infty$,*

$$E_{f_0} \Pi \left[\{ f \,:\, ||f - f_0||_2^2 > \mathcal{L}_n n^{-2\beta/(2\beta+1)} \} \,|\, X \right] \to 0,$$

where $\mathcal{L}_n = \log^\delta n$ for some $\delta > 0$. The same conclusion holds with $\mathcal{L}_n = \log^{2+\delta} n$ instead if the prior Π is replaced by the truncated Horseshoe prior (6) *with uniform scaling $\tau = n^{-4}$*

Theorem 1 establishes that, for the OT and truncated Horseshoe priors, the posterior 'in probability' contraction rate is nearly minimax (up to logarithmic factors) over Sobolev balls $\mathcal{S}^\beta(F)$. Since nowhere in the definition of the prior the knowledge of β is required, Theorem 1 shows that these priors are adaptive to the Sobolev smoothness of f_0 (in a minimax sense). This result can be compared to Theorem 1 in [2], where the same upper bound is obtained for OT priors with bounded densities and finite second-order moment. The key difference arises from our proof technique: rather than applying Markov's inequality working under

the posterior expectation, we work directly at the level of the posterior probability. This approach allows us replace the finite second moment condition by the weaker tail condition (H3), enabling the use of heavier-tailed priors, which may have unbounded expectation, such as the Cauchy and Horseshoe.

2.2 Lower bound for polynomial decay sequence

A natural choice of prior to recover $f_0 \in \mathcal{S}^\beta(F)$ is the HT(α) prior, which has polynomially decaying coefficients

$$\text{for all } k \geq 1, \qquad \sigma_k = k^{-1/2-\alpha}. \tag{8}$$

In [2], it is shown that this prior leads to adaptation in the over-smoothing case, i.e., if one takes $\alpha > \beta$, the posterior contraction rate is the minimax rate over $\mathcal{S}^\beta(F)$. The next result provides a matching lower bound for the posterior contraction rate in the under-smoothing case $\alpha < \beta$.

Theorem 2. *Let $f_0 \in \mathcal{S}^\beta(F)$ and Π be the HT(α) series prior with $1/2 < \alpha < \beta$. For a large enough positive constant $\delta > 0$, define*

$$\xi_n := n^{-\alpha/(2\alpha+1)} \log^{-\delta} n.$$

Then as $n \to \infty$,

$$E_{f_0} \Pi \left[\{ f \,:\, \|f - f_0\|_2 \leq \xi_n \,|\, X \right] \to 0.$$

Under the lesser assumption $0 < \alpha < \beta$, the following in-expectation bound holds, as $n \to \infty$,

$$E_{f_0} \int \|f - f_0\|_2^2 \, d\Pi(f \,|\, X) \gtrsim n^{-2\alpha/(2\alpha+1)}.$$

Theorem 2 shows that in the under-smoothing case $\alpha < \beta$, the posterior cannot concentrate at a rate faster than $n^{-\alpha/(2\alpha+1)}$, which is slower than the minimax rate over $\mathcal{S}^\beta(F)$. This demonstrates that the adaptive property of the HT(α) prior is fundamentally one-sided, meaning that α must be chosen large enough to ensure adaptation. This fact provides a strong incentive to use the OT prior (4), which, according to Theorem 1, always guarantees adaptation by ensuring that one is 'always in the over-smoothing case' (hence the name).

3 Inhomogeneous smoothness and sparse Besov rates

Consider now a function $f_0 \in L^2$ with wavelets coefficients $\{f_{jk}, j \geq -1, 1 \leq k \leq 2^j - 1\}$ in a regular orthonormal wavelet system of $L^2[0,1]$ (e.g. a boundary corrected Daubechies system see [4] or [7] for a construction). The observation sequence is now given by

$$X_{jk} | f_{0,jk} \sim \mathcal{N}(f_{0,jk}, 1/n). \tag{9}$$

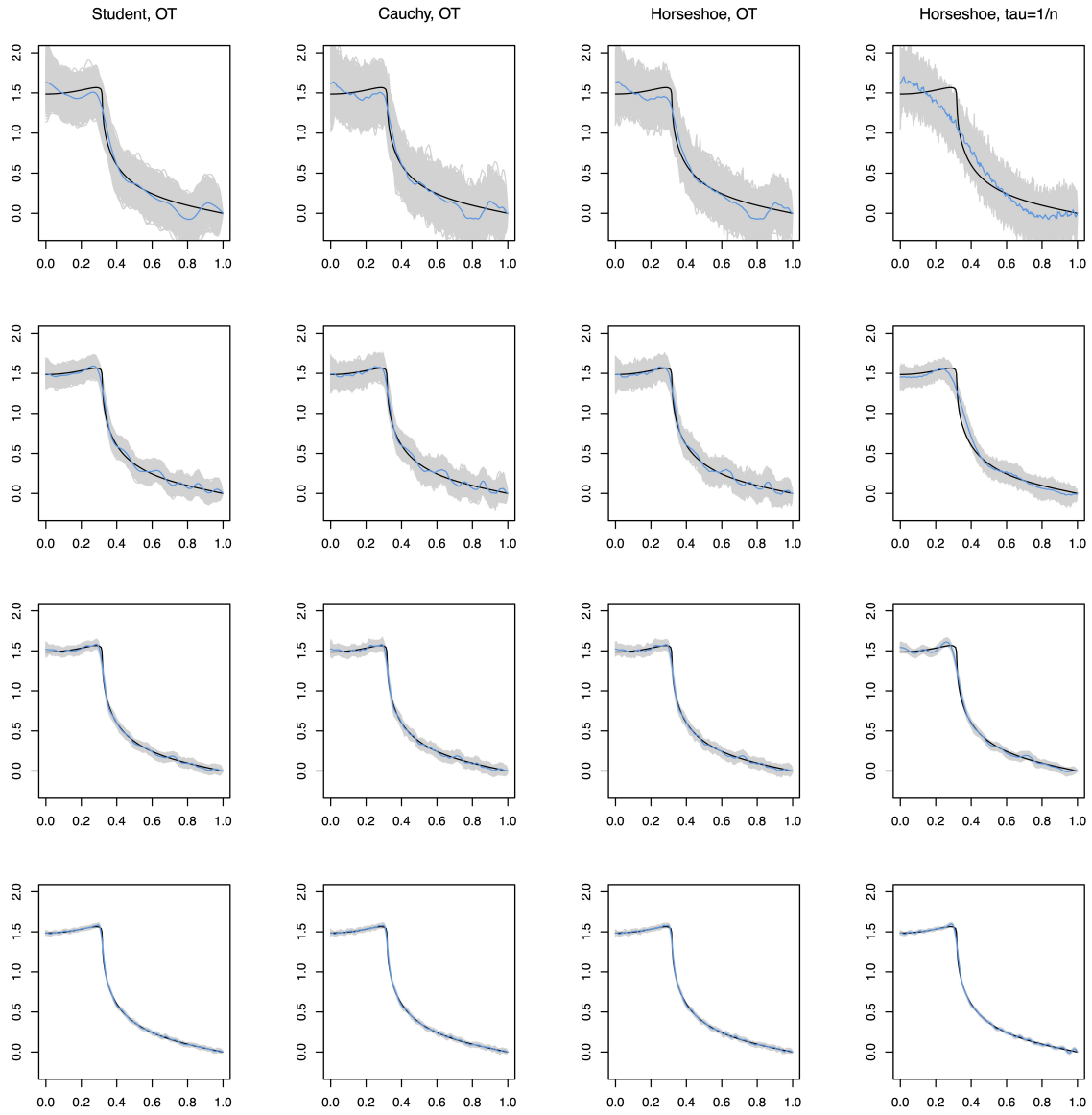


Figure 1: White noise model: true function (black), posterior means (blue), 95% credible regions (grey), for $n = 10^3, 10^4, 10^5, 10^6$ top to bottom and for the four considered priors left to right.

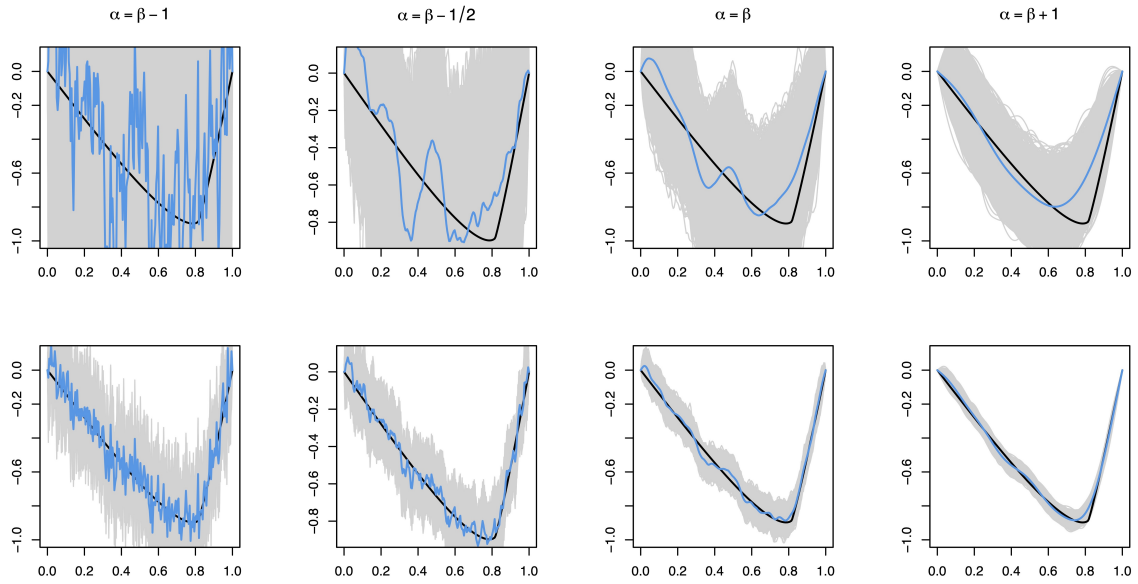


Figure 2: White noise model: true function (black), posterior means (blue), 95% credible regions (grey), for $n = 2 \times 10^2$ (top row) and $n = 2 \times 10^4$ (bottom row). Student HT(α) prior with $\nu = 3$ degrees of freedom, for $\alpha = 0.75, 1.25, 1.75, 2.75$ left to right, where the Sobolev regularity of the truth β is (almost) 1.75.

Given $p \geq 1$ write $||f_{j\cdot}||_p := (\sum_k |f_{jk}|^p)^{1/p}$ for the L^p norm of the sequence $(f_{jk})_k$ with $k \in \{0, \dots, 2^j - 1\}$ the true regression function f_0 is now assumed to live in a Besov-type ball $B^s_{p,q}(F)$ defined for $s, F > 0$ and $p, q \geq 1$ by

$$B^s_{pq}(F) = B^s_{pq}([0,1], F) := \{f \in L^p([0,1]) \,:\, ||f||_{B^s_{pq}} \leq F\}, \tag{10}$$

with

$$||f||_{B^s_{pq}} := \left(\sum_{j \geq -1} 2^{qj(s+1/2-1/p)} ||f_{j\cdot}||_p^q \right)^{1/q}. \tag{11}$$

Whenever $s' := s - 1/p + 1/p' > 0$, we have $B^s_{pq} \subset B^{s'}_{p'q} \subset L^{p'}$, therefore we can try and estimate a function $f_0 \in B^s_{pq}$ in $L^{p'}$-norm from the sequence model (9). The difficulty of this estimation problem is highly dependent on the loss parameter p' and is notoriously difficult whenever p' is large. This can be seen through the minimax rate n^{-r}(up to log factors) where

$$r := \begin{cases} s/(1+2s), & \text{if } p' < (2s+1)p, \\ s'/(1+2(s-1/p)), & \text{if } p' \geq (2s+1)p. \end{cases} \tag{12}$$

The reason for the elbow in the optimal rate when $p' \geq (2s+1)p$ can be explained by the fact that for non-homogeneous functions in B^s_{pq} the $L^{p'}$ norm can be heavily modified with only a few changes in the coefficients of the functions. Making the statistical problem of function reconstruction in p' norm from noisy coefficients way harder in this region. Regarding

optimality of rates, historically the first systematic study of non-matched situations where the parameter of the loss function does not match the norm index defining the functional class is due to Nemirovskii [12] and to Nemirovskii, Polyak and Tsybakov [10, 11], where in particular it was noticed that linear estimators can be suboptimal. Donoho and Johnstone [5] studied this type of questions in sequence space, while Donoho, Johnstone, Kerkyacharian and Picard [6] investigate getting *adaptive* minimax rates in density estimation for Besov classes. For such classes, it turns out that the information-theoretic boundaries can be described by three 'zones', depending on the loss function and the characteristics of the Besov space – we refer to [8] for an in-depth discussion –: a *regular* zone, where best linear estimators achieve the minimax rate and the rate is the classical $n^{-s/(2s+1)}$, an *intermediate* zone, where the rate is still the same, but linear estimators are suboptimal, and a *sparse* zone, were the rate changes and can be expressed as a combination of the regularity and loss function parameters. This illustrates the delicate interplay between measures of loss and regularity, which in particular leads to significant, polynomial–in–n differences in convergence rates. Aside of wavelet methods, an adaptive local bandwidth selector based on Lepski's method was developed for kernel–based estimators in [9]. Coming back to Bayesian methods, Gaussian Processes can be suboptimal when estimating functions with inhomogeneous smoothness (a fact formally proved in [1] in the white noise model, with mean square loss); the following result shows that in contrast the OT prior achieves near-minimax adaptive rates

Theorem 3. *Let $p, q \in [1, \infty]$, $s, F > 0$ and $f_0 \in B^s_{pq}(F)$. Consider observations from model (9), let $p' \in [1, \infty]$ and assume $s - 1/p + 1/(p' \vee 2) > 0$. Let Π be the OT wavelet series prior sampling coefficients f_{jk} independently as*

$$f_{jk} = 2^{-j^2} \zeta_{jk},$$

where (ζ_{jk}) are i.i.d. copies of a heavy-tailed random variable ζ satisfying conditions H1–H3 (and the added moment condition $E|\zeta| < \infty$ when $p' = \infty$). Then as $n \to \infty$,

$$E_{f_0} \Pi \left[\{ f \,:\, \|f - f_0\|_{p'} \geq \mathcal{L}_n n^{-r} \} \,|\, X \right] \to 0,$$

where $\mathcal{L}_n = \log^{\delta} n$ for some $\delta > 0$ and r is given by (12).

References

[1] S. Agapiou and S. Wang. "Laplace priors and spatial inhomogeneity in Bayesian inverse problems". In: *Bernoulli* 30.2 (2024), pp. 878–910. ISSN: 1350-7265,1573-9759. DOI: 10.3150/22-bej1563. URL: https://doi.org/10.3150/22-bej1563.

[2] Sergios Agapiou and Ismaël Castillo. "Heavy-tailed Bayesian nonparametric adaptation". In: *Ann. Statist.* 52.4 (2024), pp. 1433–1459. ISSN: 0090-5364,2168-8966. DOI: 10.1214/24-aos2397. URL: https://doi.org/10.1214/24-aos2397.

[3] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. "The horseshoe estimator for sparse signals". In: *Biometrika* 97.2 (2010), pp. 465–480. ISSN: 0006-3444,1464-3510. DOI: 10.1093/biomet/asq017. URL: https://doi.org/10.1093/biomet/asq017.

[4] Albert Cohen, Ingrid Daubechies, and Pierre Vial. "Wavelets on the Interval and Fast Wavelet Transforms". In: *Applied and Computational Harmonic Analysis* 1.1 (1993), pp. 54–81. ISSN: 1063-5203. DOI: https://doi.org/10.1006/acha.1993.1005. URL: https://www.sciencedirect.com/science/article/pii/S1063520383710055.

[5] David L. Donoho and Iain M. Johnstone. "Ideal spatial adaptation by wavelet shrinkage". In: *Biometrika* 81.3 (1994), pp. 425–455. ISSN: 0006-3444,1464-3510. DOI: 10.1093/biomet/81.3.425. URL: https://doi.org/10.1093/biomet/81.3.425.

[6] David L. Donoho et al. "Density estimation by wavelet thresholding". In: *Ann. Statist.* 24.2 (1996), pp. 508–539. ISSN: 0090-5364,2168-8966. DOI: 10.1214/aos/1032894451. URL: https://doi.org/10.1214/aos/1032894451.

[7] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Vol. 40. Cambridge university press, 2015.

[8] Wolfgang Härdle et al. *Wavelets, approximation, and statistical applications*. Vol. 129. Lecture Notes in Statistics. Springer-Verlag, New York, 1998, pp. xviii+265. ISBN: 0-387-98453-4. DOI: 10.1007/978-1-4612-2222-4. URL: https://doi.org/10.1007/978-1-4612-2222-4.

[9] O. V. Lepski, E. Mammen, and V. G. Spokoiny. "Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors". In: *Ann. Statist.* 25.3 (1997), pp. 929–947. ISSN: 0090-5364,2168-8966. DOI: 10.1214/aos/1069362731. URL: https://doi.org/10.1214/aos/1069362731.

[10] A. S. Nemirovskiĭ, B. T. Polyak, and A. B. Tsybakov. "Estimates of the maximum likelihood type for a nonparametric regression". In: *Dokl. Akad. Nauk SSSR* 273.6 (1983), pp. 1310–1314. ISSN: 0002-3264.

[11] A. S. Nemirovskiĭ, B. T. Polyak, and A. B. Tsybakov. "The rate of convergence of nonparametric estimates of maximum likelihood type". In: *Problemy Peredachi Informatsii* 21.4 (1985), pp. 17–33. ISSN: 0555-2923.

[12] A. S. Nemirovskiy. "Nonparametric estimation of smooth regression functions". In: *Soviet J. Comput. Systems Sci.* 23.6 (1985), pp. 1–11. ISSN: 0882-4002.

MCD: Marginal Contrastive Discrimination

Katia Meziani [1] & Aminata Ndiaye [2] & Benjamin Riu [3]

[1] *Université Paris Dauphine PSL, France meziani@ceremade.dauphine.fr*
[2] *Université Paris Dauphine PSL, France aminata.ndiaye@dauphine.eu*
[3] *Ecole Polytechnique, Université Paris Dauphine PSL, France*

Résumé. L'estimation de densité conditionnelle CDE (conditional density estimation) constitue un enjeu central en statistique et en apprentissage machine qui s'avère particulièrement complexe lorsque la dimension de l'ensemble de conditionnement est grande. Pour répondre à cette problématique, nous proposons une méthode innovante inspirée des méthodes de contraste bruité (Noise-contrastive methods). Cette approche reformule le problème d'estimation de densité conditionnelle en deux sous-problèmes plus simples : l'estimation de densité et la classification binaire. Notre méthode, appelée Marginal Contrastive Discrimination (MCD), présente des performances comparables, voire supérieures, à celles des approches existantes de l'état de l'art, en particulier dans des contextes où la dimension de l'ensemble de conditionnement est grande.

Mots-clés. Estimation de densité conditionnelle, méthode non paramétrique,apprentissage supervisé, méthodes de contraste bruité.

Abstract. Conditional density estimation is a central problem in statistics and machine learning, particularly challenging when the dimensionality of the conditioning set is high. To address this issue, we propose an innovative method inspired by noise-contrastive methods. This approach reformulates the conditional density estimation problem into two simpler sub-problems: density estimation and binary classification. Our method, called Marginal Contrastive Discrimination (MCD), demonstrates performance that is comparable to, and sometimes surpasses, state-of-the-art conditional density estimation approaches, especially in scenarios where the dimensionality of the conditioning set is high.

Keywords. Conditional density estimation, non-parametric method, supervised learning, noise-contrastive methods

1 Introduction

1.1 Motivation

We consider the problem of conditional density estimation (**CDE**), which is a major topic of interest in the fields of statistics and machine learning..
We consider a couple of random variables (X, Y) with $\mathcal{X} \times \mathcal{Y} \subseteq R^p \times R$. We denote by f_X , f_Y and $f_{X,Y}$ the densities of X, Y and (X, Y) with respect to a dominant measure. Our goal is to estimate the conditional density function:

$$\begin{aligned} \mathcal{X} \times \mathcal{Y} &\longrightarrow R_+ \\ x, y &\longmapsto f_{Y \mid X=x}(y) \end{aligned}$$

This problem is at the root of the majority of machine learning tasks. Supervised learning techniques aims at estimating the conditional mean. In the binary classification setting, these two tasks are equivalent, since $E[Y \mid X = x] = P[Y = 1 \mid X = x]$. In the regression setting where Y is a continuous variable, the conditional density is far more informative than the mean value. This is especially true when the conditional distribution is multi-modal, heteroscedastic or heavy tailed.

1.2 From noise contrastive methods to conditional density estimation (CDE)

Historically, the first attempts were based on the use of Bayes' formula which transforms the estimation of the conditional density into the estimation of two density functions, thus allowing the use of techniques dedicated to density estimation.

$$f_{Y \mid X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)},\ f_X \neq 0$$

A major flaw of reformulating a conditional density as the ratio of two densities, is that even if Y is low dimensional, we will still incur the curse of dimensionality if X is high-dimensional.

The noise-contrastive estimation (Gutmann and Hyvärinen [2010]) is a family of techniques for density estimation that reformulates the problem as binary classification. A known distribution g is introduced, and synthetic samples x' are drawn from

it. These synthetic samples are combined with the original dataset, and a binary target Z. Z is assigned to each observation, it takes the value 1 for original data and the value 0 otherwise. A binary classifier is then trained to predict this target value conditionally on the associated observation. The conditional probability estimated corresponds to the **contrast** q, which can be directly linked to the density of the original observations: $q(x) := P[Z = 1 \,|\, X = x] = \frac{f_X(x)}{f_X(x)+g(x)}$ and $f_X(x) = g(x)\frac{q(x)}{1-q(x)}$. Under certain conditions on g, we can obtain an estimation of the density function.

2 Marginal Contrastive Discrimination (MCD)

Our method, **MCD** , is grounded on the noise contrastive method approach but we reformulate the problem differently, focusing on the contrast between the joint law $f_{X,Y}$ compared to the marginal laws f_X and f_Y .

2.1 Contrast function

Definition 1 *[**Marginal Contrast function MCF**(r)] Let r be a real such that $r \in (0,1)$. Consider (X,Y) a couple of random variables taking values in $\mathcal{X} \times \mathcal{Y}$. The Marginal Contrast Function $q(\cdot,\cdot)$ of couple (X,Y), with ratio r, is defined as:*

$$\begin{aligned} \mathcal{X} \times \mathcal{Y} &\longrightarrow [0,1) \\ (x,y) &\longmapsto q(x,y) := \frac{r\, f_{X,Y}(x,y)}{r\, f_{X,Y}(x,y) + (1-r)\, f_X(x)\, f_Y(y)} \end{aligned}$$

The introduction of this function is motivated by the fact that it will enable us to transform a hard conditional density estimation problem into an easier supervised learning estimation. This translates to Fact 1 provided below:

Fact 1 *Let r be a real such that $r \in (0,1)$. Consider (X,Y) a couple of random variables taking values in $\mathcal{X} \times \mathcal{Y}$. For all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, we have*

$$f_{Y \,|\, X=x}(y) = f_Y(y) \frac{q(x,y)}{1-q(x,y)} \frac{1-r}{r}$$

*where $q(\cdot,\cdot)$ denotes the **MCF**(r).*

This means estimating the conditional density is equivalent to estimating both the marginal density of Y and the marginal contrast function q.

2.2 Marginal discrimination conditions

To transform the problem of estimating q into a problem of supervised learning, we need to introduce a couple of random variables (W, Z) such that Z corresponds to the binary class associated to the observation W.First, we will characterise (W, Z) by introducing the *Marginal Discrimination Conditions* of the couple (X, Y):

Definition 2 ***[Marginal Discrimination Condition MDC**(r)**]***

*Let r be a real such that $r \in (0, 1)$. Consider (X, Y) two random variables taking values in $\mathcal{X} \times \mathcal{Y}$. A couple of random variables (W, Z) satisfies the Marginal Discrimination Condition (**MDC**) of the couple (X, Y) with ratio r if*

C.1 *The random variable Z follows a Bernoulli law of parameter r: $Z \sim \mathcal{B}(r)$.*

C.2 *The support of W is $\mathcal{W} = \mathcal{X} \times \mathcal{Y}$.*

C.3 $\forall (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \{0; 1\},\ f_{W \mid Z=z}(x, y) = 1_{z=1} f_{X,Y}(x, y) + 1_{z=0} f_X(x) f_Y(y)$.

2.3 Estimation of the contrast function through supervised learning

The marginal contrast function q can be estimated though a supervised learning task, using either a regressor or a binary classifier, provided we have access a sample of random variables which satisfy the $\mathbf{MDC}(r)$, as announced in the following proposition:

Proposition 1 ***[Contrast Estimation]***

Let r be a real such that $r \in (0, 1)$. Consider (X, Y) a couple of random variables taking value in $\mathcal{X} \times \mathcal{Y}$. Denote by q the marginal contrast function of couple (X, Y) with ratio r. We have:

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y},\ q(x, y) = E[Z \mid W = (x, y)] = P[Z = 1 \mid W = (x, y)]$$

It is possible to construct a training set (independent and identically distributed or just identically distributed) which satisfies the $\mathbf{MDC}(r)$.

3 Experiments

We evaluated our model in the high-dimensional setting ($n = 1000$, $p = 15000$) alongside other density models: **MDN** (Bishop [1994]), (**N.Flow**) (Rezende and Mohamed [2016]) , **RFCDE** (Pospisil and Lee [2018]), **NNKCDE** (Pospisil [2020]), **Deepcde**) (D'Isanto and Polsterer [2018]) and **FlexCode** (Izbicki and Lee [2017]). The results are presented in table 1.The **MCD** method consistently demonstrates strong performance compared to other state-of-the-art methods. The only density for which **MCD** doesn't achieve the best result is the **Gaussian Mixture** density model. However, even in this case, **MCD** is very close to the best performance (0.0014 for **MDN** vs 0.0018 for **MCD:MLP**).

Density	**MCD:MLP**	**MCD:SVM**	**NNKCDE**	**N.Flow**
GaussianMixture	0.0018±0.0002	0.0042±0.0	0.0042±0.0	0.0015±0.0002
LinearGaussian	**0.0751±0.004**	0.076±0.0001	3.9774±0.0	0.1753±0.0246
LinearStudentT	**0.0318±0.0005**	0.0321±0.0	1.877±0.0	1.392±0.241

Density	**MDN**	**Deepcde**	**RFCDE**	**FlexCode:N.N.**
GaussianMixture	**0.0014±0.0002**	0.4296±0.0838	0.0362±0.0005	0.0291±0.0065
LinearGaussian	0.291±0.0344	91.2839±16.8164	14.8472±0.0002	4.6657±1.2499
LinearStudentT	1.3642±0.1322	16.9251±5.8563	2.7034±0.0002	1.0618±0.1308

Table 1: Kullback-Leibler divergence (mean ± std over 50 replications) for benchmark methods and the MCD method with different classifiers across 3 density models. For each density, the result of the best method is highlighted in bold and the second-best result is underlined. We evaluated 3 classifiers with the MCD method: **MLP** a multilayer perceptron and **SVM**.

4 Conclusion

We considered the problem of conditional density estimation. We introduced a new method inspired by contrastive learning. **MCD** reformulates the initial task into a problem of supervised learning. We evaluated our method on a benchmark of density models and compared it with state-of-the-art methods. **MCD** achieved impressive performance.

References

Christopher M. Bishop. Mixture density networks. Technical report, 1994.

A. D'Isanto and K. L. Polsterer. Photometric redshift estimation via deep learning - Generalized and pre-classification-less, image based, fully probabilistic redshifts. Astronomy & Astrophysics, 609:A111, January 2018. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201731326. URL https://www.aanda.org/articles/aa/abs/2018/01/aa31326-17/aa31326-17.html. Publisher: EDP Sciences.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pages 297–304. JMLR Workshop and Conference Proceedings, March 2010. URL https://proceedings.mlr.press/v9/gutmann10a.html. ISSN: 1938-7228.

Rafael Izbicki and Ann B. Lee. Converting high-dimensional regression to high-dimensional conditional density estimation. Electronic Journal of Statistics, 11(2):2800–2831, January 2017. ISSN 1935-7524, 1935-7524. doi: 10.1214/17-EJS1302. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-11/issue
Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Taylor Pospisil. NNKCDE: Nearest Neighbor Conditional Density Estimation, June 2020. URL https://github.com/tpospisi/NNKCDE. original-date: 2018-02-19T15:10:04Z.

Taylor Pospisil and Ann B. Lee. RFCDE: Random Forests for Conditional Density Estimation. arXiv:1804.05753 [cs, stat], May 2018. URL http://arxiv.org/abs/1804.05753. arXiv: 1804.05753.

Danilo Jimenez Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. arXiv:1505.05770 [cs, stat], June 2016. URL http://arxiv.org/abs/1505.05770. arXiv: 1505.05770.

Nonparametric estimation for an integral transform of the transition density

Felix Schürzinger [1]

[1] *Université Paris Cité, CNRS, MAP5, F-75006 Paris, France. felix.schurzinger@u-paris.fr*

Abstract. We consider observation of N independent stochastic processes $(X_s^{(i)})_{s\in[0,2T]}$, $i = 1, ..., N$ on a fixed time interval $[0, 2T]$ that solve a stochastic differential equation $dX_s = b(X_s)ds + \sigma(X_s)dW_s, X_0 = x_0 \in \mathbb{R}$ where $W = (W_s)_{s\in[0,2T]}$ is a standard Brownian motion, and $b, \sigma \in C^1(\mathbb{R})$. Our aim is to construct a nonparametric least-squares projection estimator for the function $F_t(x) := \int_{-\infty}^{\infty} g(y)p_t(x, y)\, dy$. This is an integral transform of the transition density $p_s(x, y)$ with a known, real map $g : \mathbb{R} \to \mathbb{R}$. Note that $p_s(x, y)$ is the conditional density of X_{s+t} given $X_t = x$. Specifically, we define an estimator $\hat{F}_m$ as the minimizer of a least-squares regression contrast over a finite-(m-)dimensional subspace of $\mathbb{L}^2(A, dx)$ constructed using an orthonormal basis $(\varphi_j)_{j=0,...,m-1}$. We prove risk bounds for the estimator under various Assumptions on g and the estimation interval $A \subset \mathbb{R}$. We do this in an empirical norm $\|\cdot\|_N$ constructed using the observations $(X_s^{(i)})_{s\in[0,2T]}$, $i = 1, ..., N$ as well as an integrated norm $\|\cdot\|_f$ constructed as the standard $\mathbb{L}^2(A)$-norm weighted with the theoretical density $f(x) := \frac{1}{T}\int_0^T p_s(x_0, x)\, ds$. Next, we study the bias- and variance terms of these bounds, which allows for assessment of the rate of convergence of the estimator in the standard $\mathbb{L}^2(A)$-norm. This is done under the Assumption that the target function F_t lies in a chosen regularity space, specifically the Sobolev(-Hermite) ellipsoids. Additionally, we give a model selection procedure that determines an estimate $\hat{m}$ of the optimal dimension automatically from a random collection of models and show that the corresponding least-square estimator $\hat{F}_{\hat{m}}$ achieves the bias-variance compromise. Estimating F_t has applications in financial mathematics, such as for option pricing, wherein the known function g is a payoff function. We conclude by illustrating the performance of our estimation procedure by first using Ornstein-Uhlenbeck diffusion processes to construct our observations, along with simple candidates for g, and finally by applying it to an options pricing scenario.

Keywords. Non and semiparametric statistics, Statistics for stochastic processes

1 Definitions and assumptions

Let $X^{(1)}, ..., X^{(N)}$ be N independent solutions of the stochastic differential equation

$$dX_s = b(X_s)ds + \sigma(X_s)dW_s, \quad X_0 = x_0 \in \mathbb{R} \tag{1}$$

we solve the problem of estimating $F_t(x) = \int_{-\infty}^{\infty} g(y)p_t(x, y)\, dy$ by minimizing a least-squares contrast:

$$\gamma_N(\tau) := \frac{1}{NT}\sum_{i=1}^{N}\left[\int_0^T \tau^2(X_s^{(i)})\, ds - 2\int_0^T \tau(X_s^{(i)})g(X_{s+t}^{(i)})\, ds\right] \tag{2}$$

Defining the density $f(x) := \frac{1}{T}\int_0^T p_s(x_0, x)\, ds$ we construct a nonparametric projection estimator of F by minimizing the contrast (2). For sake of readability, we drop the indexing of F with the fixed time t. We construct a space $S_m \subset \mathbb{L}^2(A, f(x)dx)$ spanned by an $\mathbb{L}^2(A)$-orthonormal basis $(\varphi_j)_{j=0,...,m-1}$ supported on $A \subset \mathbb{R}$ such that $S_m := span\{\varphi_0, ..., \varphi_{m-1}\}$ in order to define the least-squares estimator

$$\hat{F}_m = \underset{\tau \in S_m}{\operatorname{argmin}}\, \gamma_N(\tau) = \sum_{j=0}^{m-1} (\hat{\theta}_m)_j \varphi_j \in S_m \tag{3}$$

Solving the minimization problem, we obtain that the vector $\hat{\theta}_m$ with coefficients $(\hat{\theta}_m)_j$ is equal to $\hat{\theta}_m = \hat{\Psi}_m^{-1}\hat{Z}_m$ with the matrix and the vector:

$$\hat{\Psi}_m := \left(\frac{1}{NT} \sum_{i=1}^{N} \int_0^T \varphi_j(X_s^{(i)}) \varphi_{j'}(X_s^{(i)})\, ds \right)_{j,j' \in \{0,...,m-1\}} \in \mathbb{R}^{m \times m}$$

$$\hat{Z}_m := \left(\frac{1}{NT} \sum_{i=1}^{N} \int_0^T \varphi_j(X_s^{(i)}) g(X_{s+t}^{(i)})\, ds \right)_{j \in \{0,...,m-1\}} \in \mathbb{R}^{m}$$

We denote the theoretical counterpart to $\hat{\Psi}_m$ as $\Psi_m := \mathbb{E}\left[\hat{\Psi}_m\right]$. To ensure our estimator is well-defined, we need the matrix $\hat{\Psi}_m$ to be invertible. This is fulfilled under conditions on the basis as outlined in Lemma 1 of Comte and Genon-Catalot (2020). In that case, our estimator is the unique solution of the optimization problem of minimizing the contrast (2). We may now define for real, square integrable maps τ_1, τ_2 :

- An empirical scalar product $\langle \tau_1, \tau_2 \rangle_N := \frac{1}{NT}\sum_{i=1}^N \int_0^T \tau_1(X_s^{(i)})\tau_2(X_s^{(i)})\, ds$ and the corresponding empirical norm $\|\tau_1\|_N^2 := \frac{1}{NT}\sum_{i=1}^N \int_0^T \tau_1^2(X_s^{(i)})\, ds$
- A theoretical scalar product $\langle \tau_1, \tau_2 \rangle_f := \mathbb{E}\left[\langle \tau_1, \tau_2 \rangle_N\right] = \int_{-\infty}^{\infty} \tau_1(x)\tau_2(x) f(x)\, dx$ and the corresponding integrated norm $\|\cdot\|_f$

Throughout our paper, we require some basic Assumptions to be fulfilled.

Firstly, we need that $b, \sigma \in C^1(\mathbb{R})$ with bounded derivatives b' and σ' and that σ is upper- and lower bounded by positive constants (**Assumption 2.1**). This ensures that the transition density p_t is well-defined. It also allows us to derive upper bounds on the transition density $p_t(x, y)$ as in Proposition 3.1 from Comte and Marie (2024) and ensures that the density f is well-defined and upper bounded on $\mathbb{R}$.

Next, we define for our orthonormal basis $(\varphi_j)_{j=0,...,m-1}$ with support $A \subset \mathbb{R}$ the quantity $L(m) := 1 \vee \left(\sup_{x \in A} \sum_{j=0}^{m-1} \varphi_j(x)^2\right)$. We assume the considered dimensions $m \in \mathbb{N}$ of our space of estimation S_m satisfy a stability condition, meaning there exists a $q \in \mathbb{N}$ such that

$$L(m)\left(\|\Psi_m^{-1}\|_{op} \vee 1\right) \leq \frac{\mathfrak{c}_\Lambda}{2} \frac{NT}{\log(NT)}$$

where $\mathfrak{c}_\Lambda = \frac{1-\log(2)}{(1+p)T}$ and $p = 2(q+4)+1$ (**Assumption 2.2**). Such a stability condition is required to ensure that $\hat{\Psi}_m^{-1}$ remains well-conditioned. It ensures the stability of the least-squares estimator as underlined in Cohen et al (2013).

Lastly, we require that there exists a deterministic constant $\mathfrak{c} > 0$ and $q \in \mathbb{N}$ that do not depend on N and m such that $L(m) \leq \mathfrak{c}N^q$ (**Assumption 2.3**).

We now introduce conditions on the known function g, for which we separate two cases:

- In **Assumption 2.4**, we assume g is bounded, meaning there exists a finite $\mathfrak{c}_g := \sup_{x\in A} |g(x)| < \infty$
- In **Assumption 2.5**, we assume that g has at most polynomial growth

We note that in the second case, we rely on a compact basis support $A \subset \mathbb{R}$ to prove our risk bounds, which is avoidable under **Assumption 2.4**.

2 Risk bounds on the projection least squares estimator

We derive nonadaptive risk bounds for the projection estimator (3). We separate two cases.

Theorem 2.1 *Define $F_A := F\mathbb{1}_A$. Let **Assumptions 2.1-2.3** as well as **2.4** be fulfilled. Then, there exists a constant $\mathfrak{c}_1$ which does not depend on m or N such that for every $t \in [0,T]$:*

$$\mathbb{E}\left[\|\hat{F}_m - F_A\|_N^2\right] \leq \min_{\tau\in S_m} \|\tau - F_A\|_f^2 + \frac{2m\mathfrak{c}_g^2}{N} + \frac{\mathfrak{c}_1}{N}$$

Under **Assumption 2.5**, we first show there exists a finite constant C_g bounding $\int_{\mathbb{R}} f(x) \int_{\mathbb{R}} g(y)^2 p_t(x,y)\, dy\, dx \leq C_g$ and a lower bound $\pi_0 > 0$ such that on every compact interval $\pi_0 \leq \frac{1}{T}\int_0^T p_s(x_0,x)\, ds \leq \pi_1$ (c.f. Proposition 3.1 in Comte and Marie (2024)).

Theorem 2.2 *Given our **Assumptions 2.1-2.3** as well as **2.5**, for a compact $A \subset \mathbb{R}$, there exist constants π_0, $\mathfrak{c}_2$ and C_g independent of m, t and N such that for every $t \in [0,T]$:*

$$\mathbb{E}\left[\|\hat{F}_m - F_A\|_N^2\right] \leq \min_{\tau\in S_m} \|\tau - F_A\|_f^2 + C_g\frac{2}{\pi_0}\frac{L(m)}{N} + \frac{\mathfrak{c}_2}{N}$$

We continue by discussing risk bounds in the f-weighted norm $\|\cdot\|_f$. We first truncate our estimator to ensure its stability:

$$\tilde{F}_m := \hat{F}_m \mathbb{1}_{\Lambda_m} = \hat{F}_m \mathbb{1}_{\left\{L(m)(\|\hat{\Psi}_m^{-1}\|_{op}\vee 1) \leq c_T \frac{NT}{\log(NT)}\right\}}$$

Theorem 2.3 *Given our* ***Assumptions 2.1-2.3*** *and* ***2.4****, the truncated estimator* $\tilde{F}_m$ *satisfies:*

$$\mathbb{E}\left[\|\tilde{F}_m - F_A\|_f^2\right] \leq 9 \min_{\tau \in S_m} \|\tau - F_A\|_f^2 + 4\mathfrak{c}_1 \frac{m\mathfrak{c}_g^2}{N} + \mathfrak{c}_3 \frac{\mathfrak{c}_g^2 + \|F_A\|_f^2}{N}$$

where $\mathfrak{c}_1$ *stems from Theorem 2.1 and* $\mathfrak{c}_3$ *is a positive constant which like* $\mathfrak{c}_1$ *does not depend on* m *or* N*,* $\mathfrak{c}_g = \sup_{x \in A} |g(x)| < \infty$ *meaning also that* $\mathfrak{c}_g^2 + \|F_A\|_f^2 < \infty$ *since* f *is bounded.*

Theorem 2.4 *On a compact interval* $A \subset \mathbb{R}$ *and under* ***Assumptions 2.1-2.3*** *as well as* ***2.5****, there exist constants* π_0 *and* C_g *as well as* $\mathfrak{c}_4$ *depending only on* T *and the polynomial growth rate of* g *but not on* m *or* N *such that*

$$\mathbb{E}\left[\|\tilde{F}_m - F_A\|_f^2\right] \leq 9 \min_{\tau \in S_m} \|\tau - F_A\|_f^2 + 4\left(C_g \frac{2}{\pi_0} \frac{L(m)}{N}\right) + \frac{\mathfrak{c}_4(1 + \|F_A\|_f^2)}{N}$$

Since F_A *is a continuous function restricted to a compact interval and* f *is bounded,* $\|F_A\|_f^2 < \infty$

Finally, we discuss rates of convergence of the projection estimator for two choices of basis.

- On $A = \mathbb{R}$, we utilize the Hermite basis functions [**H**] as in Comte (2017). This sequence is an orthonormal basis of $\mathbb{L}^2(\mathbb{R}, dx)$. For all m, there exists a constant $\mathfrak{c}_h$ such that $L(m) \leq \mathfrak{c}_h \sqrt{m}$. Note that the Hermite basis functions are not orthonormal on a compact subset of $\mathbb{R}$.
- On a compact interval $A \subset \mathbb{R}$, we define for $m = 2d + 1$ odd the trigonometric basis functions [**T**] $(\varphi_j)_{j \in \mathbb{N}}$ as in Comte (2017). This is an orthonormal basis of the m-dimensional subspace $S_m \subset \mathbb{L}^2(A, dx)$ with $L(m) = m$.

To assess the rate of convergence of the bias term $\min_{\tau \in S_m} \|\tau - F_A\|_f^2$, we introduce a regularity space, beginning with a Sobolev-Hermite ellipsoid of order s and radius R:

$$W^s(A, R) := \left\{ h \in \mathbb{L}^2(A) \, : \, \sum_{k \in \mathbb{N}} a_k(h)^2 k^s \leq R \right\}$$

where $a_k(h) := \langle h, \varphi_k \rangle$ are the $\mathbb{L}^2$-coefficients of the development of $h \in \mathbb{L}^2(A)$ w.r.t. the Hermite basis [**H**] and $\langle \cdot, \cdot \rangle$ is the standard $\mathbb{L}^2(A)$-scalar product. For a bounded function g and under the Assumption that $F_A \in W^s(A, R)$, we may reach the best compromise between the bias- and variance terms by choosing a dimension $m^* := N^{\frac{1}{s+1}}$ so long as m^* fulfills **Assumption 2.2**. Then, we show that all terms in Theorem 2.3 are at most of order $N^{\frac{-s}{s+1}}$ such that in total $\mathbb{E}\left[\|\tilde{F}_m - F_A\|_f^2\right] \lesssim N^{-\frac{s}{s+1}}$

On $A \subset \mathbb{R}$ compact we fix the trigonometric basis functions [**T**] as our choice for the orthonormal basis $(\varphi_j)_j$. We introduce then for the order s and radius R the regularity space (Sobolev ellipsoid)

$$\Theta(s, R) := \left\{ h = \sum_{j=0}^{\infty} h_j \varphi_j \in \mathbb{L}^2(A) \, : \, \sum_{j=0}^{\infty} a_j^2 h_j^2 \leq R \right\}$$

where $a_j = (j-1)^s$ for j even and $a_j = j^s$ for j odd. Proposition 1.14 along with Theorem 1.9 from Tsybakov (2009) ensure then that if $F_A \in \Theta(s, R)$, we may reach the best compromise between the bias- and variance terms by choosing a dimension $m^* := N^{\frac{1}{2s+1}}$ so long as m^* fulfills **Assumption 2.2**. Then, we can show that all terms from Theorem 2.4, specifically the bias, are at most of order $N^{\frac{-2s}{2s+1}}$, and find the rate $\mathbb{E}\left[\|\tilde{F}_m - F_A\|_f^2\right] \lesssim N^{-\frac{2s}{2s+1}}$

3 Model selection

We now study a data-driven procedure to automatically select the correct dimension m. This is necessary as the values of the optimal dimensions are so far dependent on the regularity of the unknown function F. We fix the following Assumptions:

(A1) The collection of spaces S_m is nested such that $S_m \subset S_{m'}$ for $m \leq m'$

(A2) For each m, the basis $(\varphi_0, ..., \varphi_{m-1})$ of S_m satisfies that there exists a constant $c_\varphi > 0$ such that for $L(m)$ it is $L(m) \leq c_\varphi^2 m$

And consider the following collection of models

$$\widehat{\mathcal{M}}_N(\theta) := \left\{ m \in \{1, ..., NT\} \,:\, c_\varphi^2 m \left(\|\hat{\Psi}_m^{-1}\|_{op}^2 \vee 1 \right) \leq \theta \frac{NT}{\log(NT)} \right\}$$

along with its theoretical counterpart $\mathcal{M}_N(\theta)$ for a positive constant θ specified later. We limit ourselves to considering the case where g is bounded. Since our aim is to use Lemma A.4 from Comte and Marie (2024) for the proof of Theorem 3.1, we define a constant $\mathfrak{d}_T$ depending on $c_\Lambda := \frac{1-\log(2)}{(1+p)T}$, T, c_φ and the bound of the density f as in said paper. We estimate the optimal dimension as

$$\hat{m} := \underset{m \in \widehat{\mathcal{M}}_N(\mathfrak{d}_T)}{\operatorname{argmin}} \left\{ -\|\hat{F}_m\|_N^2 + \operatorname{pen}(m) \right\}$$

for the penalty $\operatorname{pen}(m) := \kappa \mathfrak{c}_g^2 \frac{m}{N}$ where κ is a numerical constant to be determined. Then, we prove an adaptive risk bound

Theorem 3.1 *Given our Assumptions, there exists constants C and C' that do not depend on m or N such that*

$$\mathbb{E}\left[\|\hat{F}_{\hat{m}} - F_A\|_N^2\right] \leq C \inf_{m \in \mathcal{M}_N(\mathfrak{d}_T)} \left(\inf_{\tau \in S_m} \|\tau - F_A\|_f^2 + pen(m) \right) + \frac{C'}{N}$$

Thus, our estimator $\hat{F}_{\hat{m}}$ automatically realizes the bias-variance tradeoff over the collection $\mathcal{M}_N(\mathfrak{d}_T)$.

4 Numerical Experiments

In the final Section of our paper, we discuss the application of our estimation procedure. First, we ensure that the theoretical guarantees translate to the case of discrete observations by connecting the discrete and the continuous approach. We study the risk $\mathbb{E}\left[\|\hat{F}_{m,\Delta} - F_A\|_N^2\right]$ for an estimator $\hat{F}_{m,\Delta}$ constructed just as in Section 1 but using discretely observed sample paths $\bar{X}^{(i)} = (\bar{X}^{(i)}_{k\Delta})_{k=0,\dots,2n}$ solving the SDE (1) with $n\Delta = T$ and $0 < \Delta < 1$. Inspired by Marie and Rosier (2023), we show that given a suitably small discretization step size Δ, the estimator $\hat{F}_{m,\Delta}$ is close in (empirical) risk to the estimator $\hat{F}_m$.

To illustrate our estimation procedure, we study how different choices for our variables affect the quality of our estimators of F given discretized observations of a chosen diffusion process. We plot adaptive least-square estimators with dimension $\hat{m}$ selected by our penalization procedure using the choices of basis outlined. We also present the application of our estimation procedure to the option pricing model of Black-Scholes as outlined in Section 5.1.3 of Lamberton and Lampeyre (2011) which in part motivated our two Assumptions on g in Section 1.

Bibliography

Cohen, A. and Davenport, M. and Leviatan, D. (2013). On the Stability and Accuracy of Least Squares Approximations. *Foundations of Computational Mathematics* 13 (5), pp. 819-834

Comte, F. (2017). *Nonparametric estimation* Spartacus-Idh.

Comte, F. and Marie, N. (2024). Nonparametric estimation of the transition density function for diffusion processes. Preprint arXiv:2404.00157

Comte, F. and Genon-Catalot, V. (2019). Regression function estimation as a partly inverse problem. *Annals of the Institute of Statistical Mathematics* 72 (4), pp. 1023 - 1054

Lamberton, D. and Lapeyre, B. (2011). *Introduction to Stochastic Calculus Applied to Finance.* CRC Press

Marie, N. and Rosier, A. (2023). Nadaraya-Watson Estimator for I.I.D. Paths of Diffusion Processes. *Scandinavian Journal of Statistics* 50 (2), pp. 589-637

Tsybakov, A. (2009). *Introduction to nonparametric estimation.* Springer

Statistique pour l'industrie

ALGORITHME GÉNÉTIQUE POUR LA CORRÉLATION DE TRANSFERT DE CHALEUR EN ÉBULLITION PAR FILM

Julien Beaucourt[1]

[1] *Autorité de Sûreté Nucléaire et de Radioprotection, France, julien.beaucourt@asnr.fr*

Résumé. Dans un réacteur nucléaire en situation accidentelle, la maîtrise de la température de la gaine du combustible constitue un enjeu essentiel pour garantir la sûreté de l'installation. L'évolution de cette température est fortement influencée par le transfert de chaleur entre la gaine et le caloporteur, en particulier en régime d'ébullition en film, où un film de vapeur isolant se forme à l'interface. Dans les modèles numériques, le coefficient de transfert de chaleur dans ce régime est traditionnellement estimé à partir de deux méthodes principales :

1. **La table de Groeneveld**, regroupant des valeurs empiriques ;
2. **Des corrélations analytiques**, telles que la corrélation de Berenson.

Nous proposons ici une méthode d'optimisation symbolique basée sur un algorithme génétique pour améliorer la construction de corrélations à partir des données tabulées de Groeneveld. Cette approche permet de générer des modèles explicites et interprétables tout en explorant efficacement l'espace des solutions possibles. La sélection des modèles optimaux sur le front de Pareto tient compte des compromis entre précision et complexité. En outre, un modèle bayésien simple est mis en œuvre pour évaluer les biais additif et multiplicatif, ainsi que l'incertitude du modèle. Les résultats obtenus soulignent le potentiel des approches symboliques pour la création de corrélations physiquement pertinentes.

Mots-clés. Algorithme génétique symbolique, machine learning, ébullition en film

Abstract. In a nuclear reactor under accidental conditions, controlling the temperature of the fuel cladding is a critical factor for ensuring reactor safety. The evolution of this temperature is strongly influenced by heat transfer between the cladding and the coolant, particularly during film boiling, where an insulating vapor film forms at the interface. In numerical models, the heat transfer coefficient in this regime is traditionally estimated using two main methods:

1. **The Groeneveld table**, which provides empirical values based on experiments.
2. **Analytical correlations**, such as the Berenson correlation.

Here we propose a symbolic optimization method based on a genetic algorithm to improve the development of correlations using tabulated data from Groeneveld. This approach generates explicit and interpretable models while efficiently exploring the space of possible solutions. Optimal models are selected based on the Pareto front, accounting for trade-offs between accuracy and complexity. Furthermore, a simple Bayesian model is implemented to assess both additive and multiplicative biases, as well as the uncertainty of the model. The results underscore the potential of symbolic approaches for creating physically relevant correlations.

Keywords. Symbolic genetic algorithm, machine learning, film boiling

1 Introduction

Dans un réacteur nucléaire à eau pressurisée (REP), le cœur du réacteur est le siège des réactions de fission produisant l'énergie nécessaire au maintien d'une température élevée dans le caloporteur (environ 300°C). L'exploitation des réacteurs est conditionnée à la démonstration de leur sûreté en situation normale ou accidentel, le confinement des matières radioactives doit être assuré, afin de garantir la protection de l'environnement et des populations.

Les accidents de perte de réfrigérant primaire (APRP), causés par une brèche sur le circuit de refroidissement, font partie des scénarios étudiés. Pendant les premières secondes du transitoire, une augmentation importante de la température de la gaine du combustible (qui constitue la première barrière) peut être observée, lorsque l'évacuation de la puissance du combustible est insuffisante. Un film de vapeur peut alors se former entre la gaine et le caloporteur, dégradant l'échange thermique et accentuant l'élévation de la température. Le coefficient d'échange thermique en ébullition en film (noté h_{film}) est un paramètre clef à ce stade (Ruyer, 2023 ; Hae Min Park, 2025).

Ce coefficient peut être déterminé à partir de corrélations issues du modèle de Berenson, initialement développé pour décrire l'ébullition en film sur une surface plane (Berenson, 1961), ou via une interpolation à partir de la table de Groeneveld (Groeneveld et col., 2003).

Dans la suite de ce document, nous présentons la structure générale de la table de Groeneveld, à partir de laquelle une corrélation analytique est proposée grâce à un modèle génétique.

2 Données d'entrée : la table de Groeneveld

La table de Groeneveld (Groeneveld, 2003) regroupe les résultats de 36 expériences (environ 30000 points). Le coefficient h_{film}, compris entre 7.10^{-3} et 64,29 kW/m²/s, est donné en fonction de quatre paramètres (la dispersion des variables est présentée sur la Figure 1) :

- Pression, notée P et comprise entre 1 et 200 bar ;
- Débit surfacique massique, noté G et entre 0 et 7000 kg/m²/s ;
- Titre thermohydraulique, noté x et compris entre -0,2 et 2 (sans dimension) ;
- Ecart entre les températures de gaine et de saturation, notée ΔT (entre 100 et 1200 K).

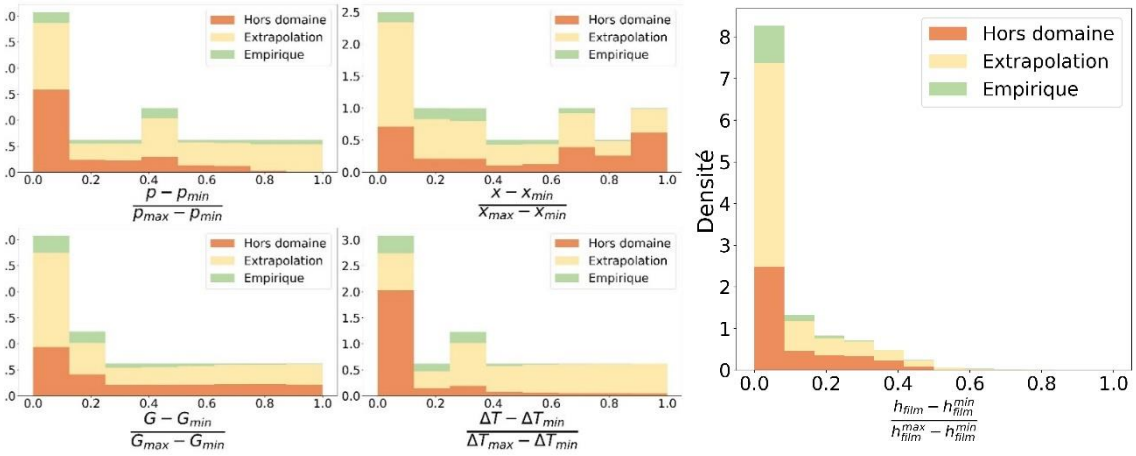


Figure 1 : Répartition des paramètres de la table de Groeneveld.

Pour éviter les difficultés liées à la prise en compte des contraintes dimensionnelles dans la construction d'une corrélation analytique, les paramètres du modèle (p, x, G et T), ainsi que le

coefficient h_{film}, sont normalisés[1]. Les données empiriques (en vert) représentent environ 10% des données de la table de Groeneveld. Les valeurs extrapolées à partir de ces données expérimentales (57% des données) apparaissent en orange. L'incertitude dans cette zone augmente à mesure que l'extrapolation s'éloigne de la plage expérimentale. Enfin, les données en rouge (environ 33% des données)), notées « Hors domaine » sur la Figure 1, correspondent à des valeurs « souvent impossibles à obtenir » (Groeneveld et col., 2003). Les scénarios pénalisants en APRP sont caractérisés par des valeurs faibles de h_{film}, toutefois, aucune restriction n'est retenue ici.

3 Modèle génétique

Un algorithme génétique (AG) est une méthode d'optimisation stochastique inspirée des processus d'évolution naturelle (Holland, 1975 ; Goldberg, 1989). Ce type d'algorithme simule la sélection naturelle en manipulant une population d'individus, où chaque individu représente une solution candidate à un problème d'optimisation donné. Des opérations de sélection, croisement et mutation (Figure 2) sont appliquées aux individus de la population, sur plusieurs générations, jusqu'à la convergence vers une solution optimale (Kramer, 2017).

Enfin, pour explorer simultanément différentes régions de l'espace des solutions et limiter le risque de convergence prématurée vers un extremum local, les AG peuvent être organisés en « modèles en îles », dans lesquels plusieurs sous-populations évoluent de manière indépendante, avec des échanges périodiques d'individus entre les îles (migration).

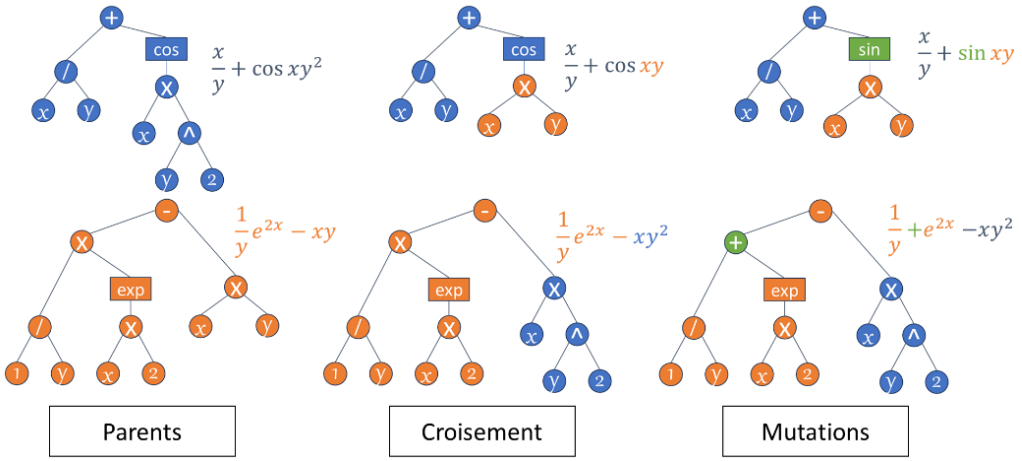


Figure 2 : Opérations de croisements et de mutations pour des arbres symboliques.

Les AG peuvent être appliqués à des problème d'optimisation symbolique (Cranmer, 2023) : les solutions candidates, analytiques, sont alors représentées par des arbres (Figure 2) dont les nœuds représentent des fonctions (addition, fonctions trigonométriques, etc.). Les nœuds terminaux correspondent à des constantes ou aux variables du problème. Ces algorithmes génétiques symboliques ont pu être utilisés pour construire des corrélations physiques à partir de données expérimentales (Jiguo Tang, 2023) : cette démarche est mise en œuvre dans ce travail afin d'établir une corrélation pour h_{film}, à partir des données de la table de Groeneveld.

A cette fin, l'algorithme développé par Miles Cranmer est utilisé (Cranmer, 2023). Ce modèle contient un grand nombre de méta-paramètres : probabilité de croisement, de mutation, fréquence et taux de migration entre les îles, etc. Dans le cadre de ce travail, les valeurs

[1] Dans la suite de cet article, les variables normalisées sont représentées par des symboles surmontés d'une barre (par exemple, $\bar{x}$ pour le titre thermohydraulique).

nominales de l'algorithme sont conservées, à l'exception des paramètres suivants :

- La complexité maximale est limitée à 30 éléments des solutions.
- Le nombre de générations est fixé à 200.
- Deux modèles sont utilisés, le premier avec 30 îles et 25 individus par îles (modèle n° 1), le second avec 60 îles et 50 individus par île (modèle n° 2).

Les fonctions de base utilisées pour la construction des solutions sont les quatre opérations arithmétiques élémentaires, l'exponentiation, le logarithme et l'exponentielle[2]. Enfin, pour accentuer l'importance des données expérimentales, les données extrapolées sont affectées d'une pondération égale à 0,1 et les données « Hors domaine » d'une pondération de 0,01.

Les deux modèles génétiques sont entrainés sur un sous-ensemble de la table de Groeneveld, correspondant à 70% des données, puis testées sur le reste des données. Les solutions sont regroupées sur le front de Pareto[3], qui met en relation la complexité des solutions et leur précision (Figure 3).

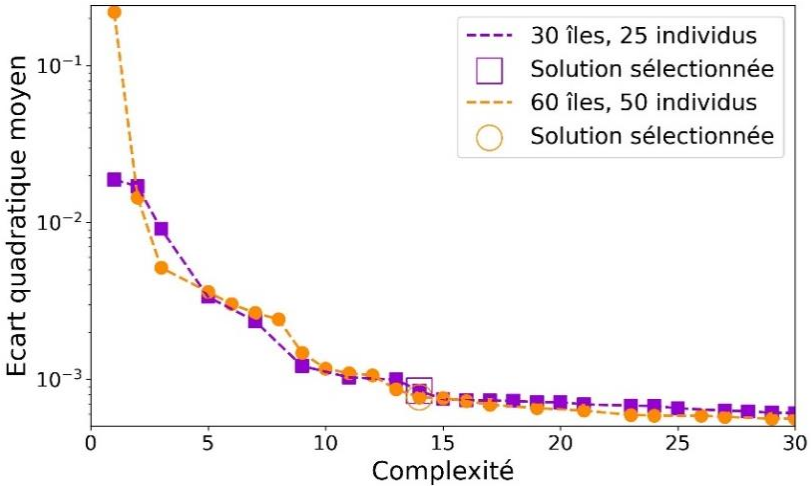


Figure 3 : Front de Pareto pour les deux modèles envisagés.

Pour chaque modèle, la solution retenue est celle qui maximise le score S (Cranmer, 2023) :

$$S = -\frac{\partial \log L}{\partial C} \tag{1}$$

avec L l'écart quadratique moyen (EQM) et C la complexité[4]. Les solutions correspondantes, pour les deux modèles envisagés, sont les suivantes :

$$\begin{cases} \bar{h}_{film}^{AG,1} = \frac{\bar{G}\bar{x}}{1{,}40-\bar{p}e^{-\bar{T}}+\bar{x}^2} \ (modèle\ n°1) \\ \bar{h}_{film}^{AG,2} = \frac{\bar{G}}{\bar{x}^2-\frac{\bar{p}e^{-\bar{T}}-1{,}25}{\bar{x}-0.04}} \ (modèle\ n°2) \end{cases} \tag{2}$$

Comme indiqué précédemment, les paramètres sont normalisés, ce qui est crucial pour garantir la cohérence dimensionnelle des solutions.

Les prédictions du modèle n° 1, plus simple, sont représentées sur la Figure 4. Les résultats sont qualitativement raisonnables : en particulier, le modèle retenu ne conduit pas à des valeurs

[2] Pour éviter le développement de solutions pathologiques, les exposants ne peuvent être que des constantes, et les fonctions logarithme et exponentielle ne peuvent pas être imbriquées (par exemple, $\log \log x$ n'est pas autorisé).
[3] Sur le front de Pareto, il est impossible de trouver une solution qui améliore la précision et la complexité.
[4] Seules les solutions dont l'EQM est inférieur à 1,5 fois celui du modèle le plus précis sont considérées.

négatives, non physiques, de h_{film} (contrairement au modèle n°2). Les écarts les plus importants sont observés pour les données extrapolées, et pour les données « hors domaine ».

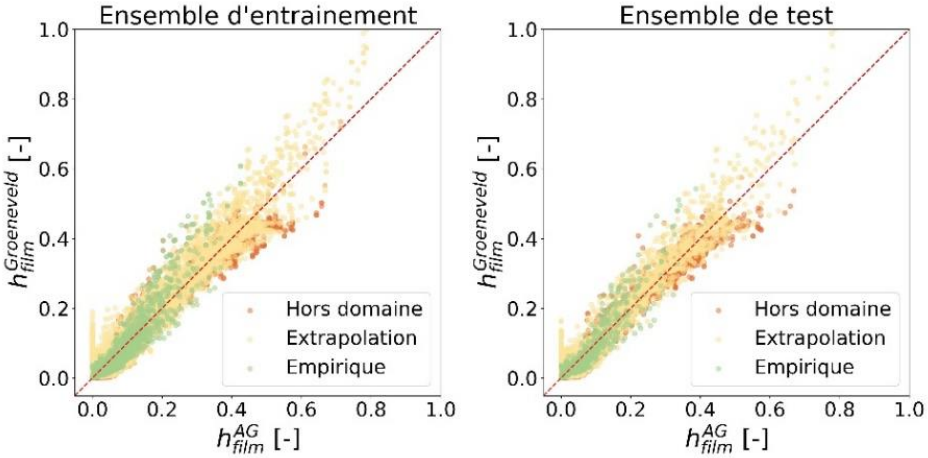


Figure 4 : Comparaison des prédictions par rapport aux données réelles (modèle n°1).

A partir de ces résultats, nous pouvons évaluer les biais multiplicatif et additif du modèle, définis respectivement comme la pente et l'intercept de la régression linéaire entre les prédictions du modèle et les valeurs de référence (sur l'ensemble de test). A cette fin, une régression linéaire bayésienne est proposée, afin de pouvoir déterminer une incertitude sur ces biais. Les distributions a priori des paramètres sont choisies de manière à limiter leur influence : plus précisément, la dispersion retenue pour les priors est importante. Pour le biais additif (intercept), une loi Beta uniforme entre 0 et 1 est prise en compte (du fait de la normalisation des données, l'intercept est nécessairement compris entre 0 et 1). Le modèle suivant est utilisé :

$$\begin{cases} \beta_0 \sim \mathrm{B}(1,1) \quad (prior\ pour\ le\ biais\ additif) \\ \beta_1 \sim \mathcal{N}(0, 10^2) \quad (prior\ pour\ le\ biais\ multiplicatif) \\ \tau \sim \Gamma(\alpha = 2, \beta = 1) \quad (prior\ pour\ la\ précision) \\ \bar{h}_{film}^{AG} | \beta_0, \beta_1, \tau \sim \mathcal{N}\left(\beta_0 + \beta_1 \cdot \bar{h}_{film}^{Groeneveld}, \frac{1}{\sqrt{\tau}}\right) \quad (vraisemblance) \end{cases} \tag{3}$$

L'échantillonnage MCMC est réalisé avec le package pyMC (Abril-Pla, 2023) sur quatre chaines, sur 800 itérations, et avec une phase de chauffe de 200 itérations. Les distributions a posteriori pour les biais sont représentées sur la Figure 5.

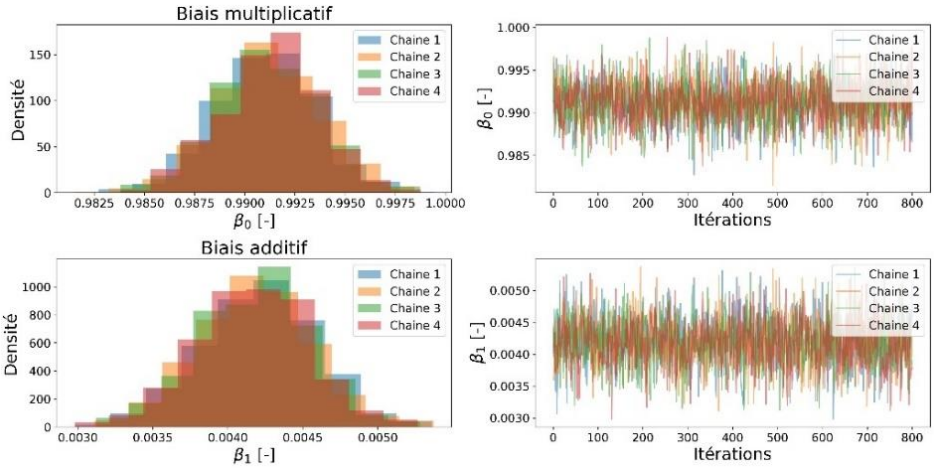


Figure 5 : Distributions a posteriori et évolution des quatre chaines MCMC.

Un biais multiplicatif de l'ordre de 0,99, et un biais additif résiduel de l'ordre de 4.10^{-3}, sont obtenus. Bien qu'encourageants, les résultats obtenus montrent qu'une dispersion importante reste présente (la valeur moyenne de $\sigma = 1/\sqrt{\tau}$ est de l'ordre de 3%). Plusieurs améliorations sont envisageables, en particulier au niveau des fonctions de base utilisées dans l'algorithme, et des méta-paramètres du modèle génétique. En outre, une restriction des intervalles de variation des paramètres d'entrée pourra être envisagée, en concertation avec les spécialistes des phénomènes d'ébullition en film, pour définir une corrélation analytique sur le domaine d'intérêt du coefficient h_{film} en situation de brèche sur le circuit de refroidissement.

4 Conclusion

L'utilisation d'un algorithme génétique symbolique (Cranmer, 2023) a permis de construire des corrélations analytiques à partir de la table de Groeneveld. Une pondération a été introduite pour accentuer l'importance des données expérimentales, et les données d'entrée sont normalisées pour garantir la cohérence dimensionnelle des solutions. Les biais additif et multiplicatif du modèle ont été estimés à partir d'un modèle de régression linéaire bayésien.

Ces résultats sont encourageants et des développements supplémentaires permettront d'optimiser les méta-paramètres du modèle. L'analyse de l'ensemble des solutions du front de Pareto permettra d'identifier les solutions les plus satisfaisantes d'un point de vue physique, en améliorant l'explicabilité du modèle. Une comparaison aux modèles analytiques de la littérature pourra alors être réalisée.

Bibliographie

Hae Min Park et col. (2025), Development of a phenomena identification and ranking table (PIRT) for intermediate break loss-of-coolant accident in PWRs, *Nuclear Engineering and Technology* 57, 103141.

P. Ruyer (2023), Recently revisited TH issues and associated R&D A French perspective, *Nuclear Engineering and Design*, 404, 112174.

P. J. Berenson (1961), *Film-Boiling heat transfer from a horizontal surface,* J. Heat Transfer, 83(3), pages 351-356.

D.C. Groeneveld et col. (2003), *A look-up table for fully developed film-boiling heat transfer*, Nuclear Engineering and Design, 225, pages 83–97.

J. H. Holland (1975), *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*, University Michigan Press.

D. E. Goldberg (1989), *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley publishing company Inc.

O. Kramer (2017), *Genetic algorithms essentials*, Springer.

M. Cranmer (2023), *Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl*, arXiv:2305.01582v3.

Jiguo Tang et col. (2023), *Development of correlations for steam condensation over a vertical tube in the presence of noncondensable gas using machine learning approach*, International Journal of Heat and Mass Transfer 201P2, 123609.

O. Abril-Pla et col. (2023), *PyMC: A Modern and Comprehensive Probabilistic Programming Framework in Python*,

Modèle d'apprentissage profond explicable pour l'amélioration de la surveillance de la radioactivité alpha aéroportée.

Roblin Arthur [1], Baccou Jean [2], Dougniaux Grégoire [3] & Velasco-Forero Santiago [4]

[1] *Autorité de Sûreté Nucléaire et de Radioprotection (ASNR), PSN-RES/SCA/LPMA, Saclay, F-91400, France. arthur.roblin@asnr.fr*
[2] *Autorité de Sûreté Nucléaire et de Radioprotection (ASNR), PSN-RES/SEMIA/LSMA, Saint-Paul-lez-Durance, F-13115, France. jean.baccou@asnr.fr*
[3] *Autorité de Sûreté Nucléaire et de Radioprotection (ASNR), PSN-RES/SCA/LPMA, Saclay, F-91400, France. gregoire.dougniaux@asnr.fr*
[4] *Mines Paris PSL, Center de Morphologie Mathématique (CMM), Fontainebleau, 77300, France. santiago.velasco@minesparis.psl.eu*

Résumé. Sur les sites nucléaires, des instruments de mesure de la radioactivité aéroportée (CAM – Continuous Air Monitor) sont déployés pour assurer la radioprotection des travailleurs. Ces instruments échantillonnent en continu les aérosols ambiants sur un filtre, mesurent la radioactivité du dépôt, et alertent l'opérateur si une contamination par un isotope transuranien est détectée. Sur les chantiers de démantèlement, où l'atmosphère tend à être particulièrement empoussiérée, Hoarau et al. (2020) ont démontré que ces instruments déclenchent de nombreuses fausses alarmes.

L'objectif du projet est de déterminer dans quelle mesure l'utilisation d'algorithmes d'intelligence artificielle permettrait d'améliorer l'analyse de la mesure nucléaire des CAM, notamment en conditions atmosphériques empoussiérées. Le principal enjeu sous-jacent est la diminution drastique du taux de fausses alarmes, sans nuire à la radioprotection des travailleurs. Au vu du caractère sensible de la problématique, on s'attache également à travailler sur l'explicabilité de l'algorithme d'IA, c'est-à-dire à la bonne compréhension du processus de décision du réseau de neurones. L'aspect "boîte noire" de la prise de décision de tels algorithmes est en effet le principal frein à leur utilisation.

La première étape a été le développement d'un code de simulation physique de spectres nucléaires, afin de construire une base de données de taille conséquente et aux paramètres variés. On a ensuite pu entraîner des modèles d'apprentissage profond avec ces données générées. L'implémentation d'une architecture de réseau de neurones convolutif a mené à de très bonnes performances de classification de spectres.

En complément, un outil intuitif de visualisation a été développé afin de comprendre, au cas par cas, ce qui motive une prédiction du modèle. On s'est également attaché à accroître l'interprétabilité du modèle. Ainsi, on s'intéresse à l'utilisation d'architectures neuronales spécifiques afin de mieux contrôler les facteurs menant au déclenchement d'une alarme.

Mots-clés. Intelligence artificielle, Deep Learning, Explicabilité, Radioactivité

Abstract. On nuclear sites, airborne radioactivity measurement instruments (CAM - Continuous Air Monitor) are deployed to ensure the radiation protection of workers. These instruments continuously sample ambient aerosols on a filter, measure the radioactivity of the deposit, and alert the operator if contamination by a transuranic isotope is detected. On dismantling sites, where the atmosphere tends to be particularly dusty, Hoarau et al. (2020) demonstrated that these instruments trigger numerous false alarms.

The aim of the project is to determine to what extent the use of artificial intelligence algorithms could improve the analysis of CAM nuclear measurements, particularly in dusty atmospheric conditions. The main underlying challenge is to drastically reduce the false alarm rate, without compromising worker radiation protection. In view of the sensitive nature of the problem, we are also working on the explicability of the AI algorithm, i.e. a good understanding of the neural network's decision-making process. The "black box" aspect of decision-making in such algorithms is the main obstacle to their use.

The first step was to develop a physical simulation code for nuclear spectra, in order to build up a large database with a wide range of parameters. Deep learning models were then trained on the generated data. The implementation of a convolutional neural network architecture led to very good spectra classification performances.

In addition, an intuitive visualization tool has been developed to understand, on a case-by-case basis, what motivates a model prediction. We also focused on enhancing the model's interpretability. We are therefore looking into the use of specific neural architectures to better control the factors leading to the triggering of an alarm.

Keywords. Artificial Intelligence, Deep Learning, Explicability, Radioactivity

1 Introduction

Sur les sites nucléaires, des instruments de mesure de la radioactivité aéroportée sont déployés afin d'assurer la radioprotection des travailleurs. Le bon fonctionnement des algorithmes traditionnellement utilisés dans ces moniteurs est conditionné à une atmosphère ambiante peu empoussiérée. Hoarau et al. (2020) ont démontré que ces conditions pouvaient ne pas être vérifiées sur les chantiers de démantèlement des centrales nucléaires, ce qui cause de nombreuses fausses alarmes. L'objectif est donc de développer un algorithme d'apprentissage statistique capable d'effectuer une classification binaire des spectres nucléaires acquis par le moniteur, et ce, pour toutes les conditions atmosphériques.

Après avoir présenté la problématique plus en détail, on présentera les modèles testés et leurs performances, avant de s'intéresser aux méthodes d'explicabilité et d'interprétabilité implémentées.

2 Problématique

La mesure nucléaire du moniteur, un spectre d'énergie, présente généralement deux pics dus à la radioactivité alpha naturelle, correspondant aux ^{214}Po et ^{218}Po - deux descendants du Radon. La figure 1 illustre la dégradation de la mesure nucléaire en conditions atmosphériques empoussiérées. Dans ces conditions, la région des transuraniens (entre 2 et 4 MeV), où apparaissent les éventuelles traces de contamination, subit une nette augmentation du bruit de fond, ce qui déstabilise les algorithmes classiques. Cette perturbation dépend de la masse d'aérosols déposés sur le filtre du moniteur, mais aussi de la granulométrie de ces aérosols.

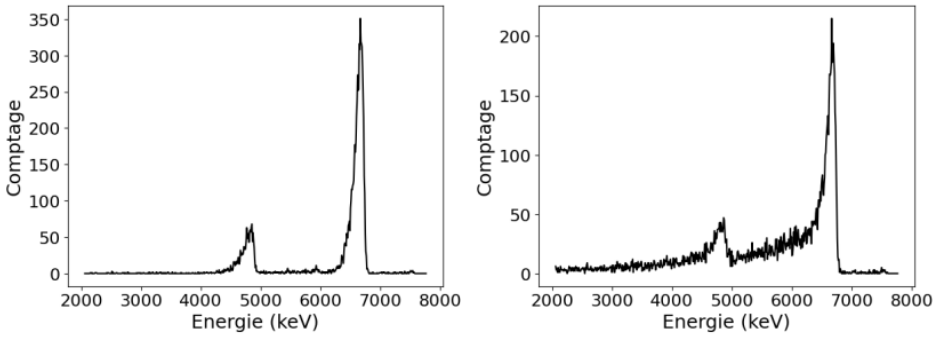


Figure 1: Spectre acquis en conditions atmosphériques normatives (à gauche) et en conditions empoussiérées (à droite).

Afin de conserver de bonnes performances de détection dans des conditions empoussiérées, on s'oriente vers la construction de modèles statistiques avancés basés sur le deep learning pour la classification binaire de spectres (présence ou absence de contamination). L'usage de ce type de modèles nécessite une base de données conséquente. La base de données expérimentale à disposition est très insuffisante, ne comptant que quelques dizaines de spectres, sans aucune trace de contamination par transuraniens. Un code de génération de spectres synthétiques a donc été préalablement développé, appuyé sur des modèles physiques établis et sur les données expérimentales. On a ainsi pu s'appuyer sur des dizaines de milliers de spectres, générés dans des conditions atmosphériques très variées, permettant une représentation de toutes les conditions atmosphériques qu'un moniteur peut être amené à rencontrer.

3 Construction et analyse du modèle de classification

3.1 Performances de plusieurs modèles

Les résultats obtenus par les différents modèles sont donnés à la table 1. L'exactitude désigne la proportion de prédictions correctes. Des modèles tels que la régression logistique ou les forêts aléatoires, après avoir mis en forme les données d'une manière inspirée d'un algorithme

traditionnel, ont permis d'améliorer les performances de classification binaire par rapport à l'algorithme actuel du moniteur. L'implémentation d'un perceptron multicouches (MLP) a encore creusé l'écart. C'est finalement un réseau de neurones convolutif (CNN) qui a obtenu les meilleures performances de classification, avec l'architecture illustrée figure 2.

Modèle	**Exactitude**	**Nombre de paramètres**
Algorithme moniteur	59,9%	1
Régression logistique	75,2%	10
Random Forest	86,6%	$\sim$ 100 000
MLP	94,6%	$\sim$ 500 000
CNN	96,8%	$\sim$ 800 000

Table 1: Performances des modèles obtenues en conditions empoussiérées, sur l'échantillon de test.

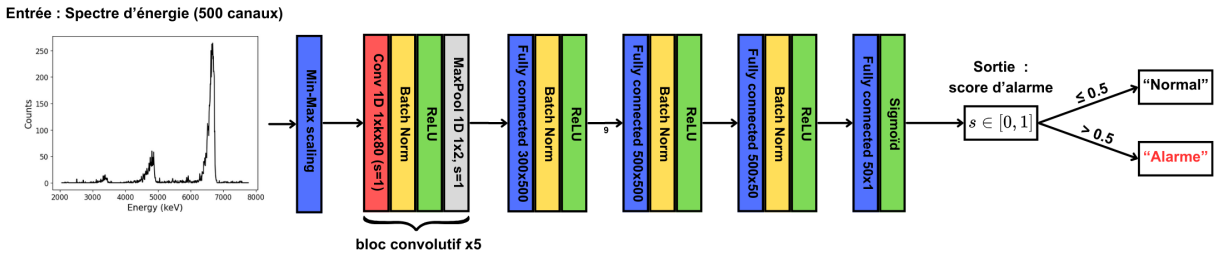


Figure 2: L'architecture du réseau convolutif. Clé de lecture : Conv 1D 1xkx30 (s=1) correspond à une couche de convolutions uni-dimensionnelles avec une largeur de noyau k, 80 filtres, un "stride" de 1 (s=1) et pas de "padding", avec k= (3,3,5,7,7) dans les cinq couches convolutives consécutives. Un Max Poolling de largeur 2 est appliqué après chaque couche de convolution.

3.2 Explicabilité du réseau de neurones

L'explicabilité d'un modèle désigne la capacité d'expliquer de façon simple et claire (souvent visuellement) pourquoi le modèle a pris une décision plutôt qu'une autre. Dans la problématique de classification de spectres, une approche particulièrement intuitive a été implémentée afin de quantifier la contribution au score de sortie de chacun des 500 canaux d'énergie des spectres. Nous présentons cette méthode dans un cadre de classification, mais son adaptation au paradigme de régression est immédiate.

Pour un spectre donné $x = (x_1, \ldots, x_{500}) \in \mathbb{N}^{500}$ labellisé $y \in \{0, 1\}$, le réseau de neurones f donne sa prédiction $\hat{y} = f(x) \in [0, 1]$. Le gradient du score de sortie du réseau vis-à-vis de chaque variable d'entrée $x_i, i \in [\![1; 500]\!]$, donne une information sur la contribution de chaque canal d'énergie du spectre à la prédiction du modèle, avec une interprétation physique directe : si la prochaine particule alpha captée par le moniteur est mesurée avec l'énergie correspondant au canal x_i, cela augmentera-t-il le score d'alarme estimé ?

On trace donc $\left(\frac{\partial f(x)}{\partial x_i}\right)_{i=1,\ldots,500}$ afin de visualiser cette information pour tout le spectre. Un exemple sur un spectre quelconque est donné à la figure 3.

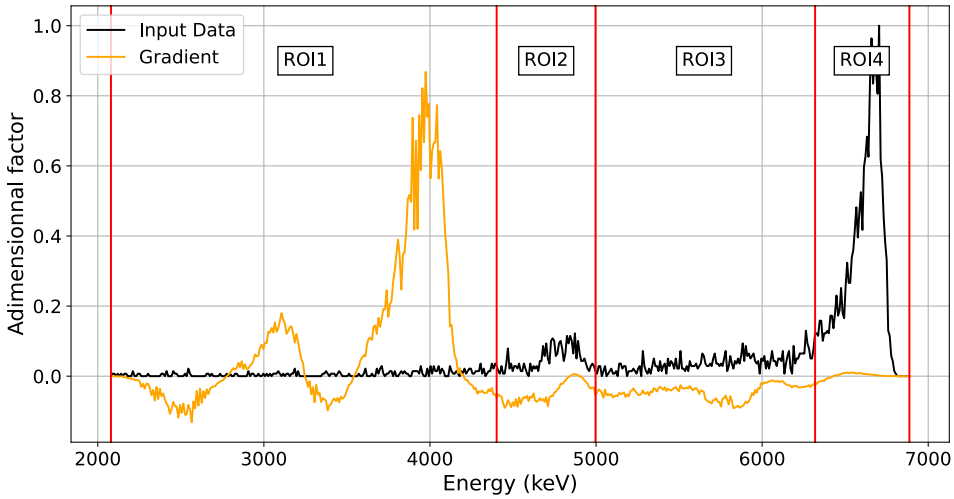


Figure 3: Gradient du score de sortie du modèle en fonction de chaque canal d'entrée. Le spectre est découpé en quatre régions d'intérêt (ROI). ROI1 correspond à la gamme d'énergie des transuraniens.

3.3 Interprétabilité du réseau de neurones

L'interprétabilité d'un modèle désigne la compréhension globale de son comportement. Les relations entre les entrées et les sorties d'un modèle sont plus transparentes. La visualisation des prises de décision, dont un exemple est visible à la figure 4, a mis en évidence un comportement importun du modèle. On observe certes un fort gradient positif dans la zone des transuraniens (ROI1), ce qui est en accord avec l'analyse physique; toutefois, certains canaux de cette zone tendent à contribuer négativement au score de sortie.

Pour assurer au modèle une meilleure interprétabilité et une robustesse accrue, il est souhaitable d'imposer à tous les coups enregistrés dans cette région de contribuer positivement à la prédiction du modèle, comme dans les algorithmes traditionnels. Pour ce faire, on a implémenté une architecture neuronale spécifique, un réseau partiellement isotonique, proposée par Kim et Lee (2024). La figure 4 fournit le gradient du score de sortie. La contribution des canaux est bien positive dans la région des transuraniens. Cependant, le gain de robustesse obtenu par cette architecture se paie par des performances de classification légèrement dégradées ($\sim$ 92% d'exactitude). Diverses modifications sont explorées afin d'améliorer cette architecture dans le cadre de notre problématique.

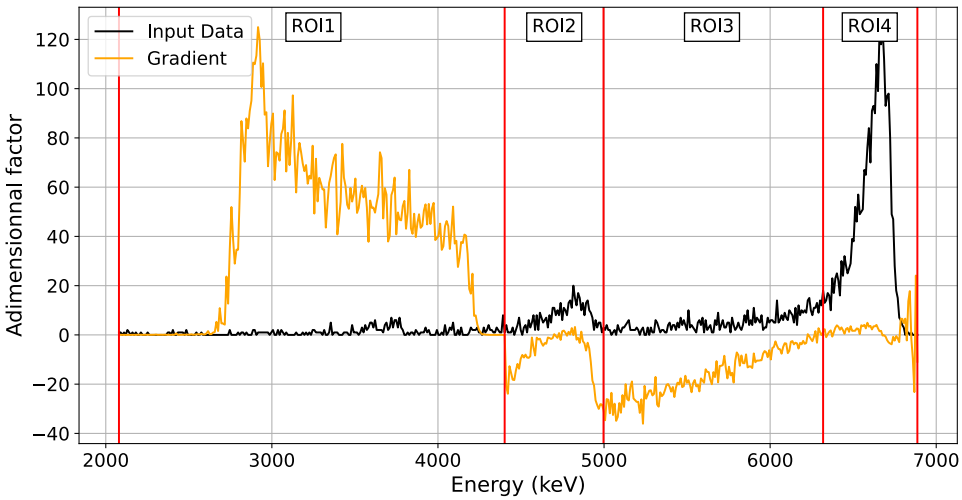


Figure 4: Gradient du score de sortie du réseau isotonique en fonction de chaque canal d'entrée. Le spectre est découpé en quatre régions d'intérêt (ROI). ROI1 correspond à la gamme d'énergie des transuraniens.

4 Conclusion

L'utilisation de modèles d'apprentissage profond a permis une amélioration drastique des performances de classification des spectres acquis en conditions empoussiérées. Le manque de transparence de la prise de décision reste un frein majeur à l'utilisation pratique des réseaux de neurones. Les approches présentées permettent d'atténuer ce problème. En complément, l'estimation des incertitudes de prédiction est un impératif dans le cadre d'une problématique aussi sensible. Dans le paradigme de classification binaire des spectres, on souhaitera associer la prédiction à un indice de confiance. Basées sur l'état de l'art de Gawlikowski et al (2023), différentes approches sont envisagées : méthodes ensemblistes, réseaux de neurones bayésiens, estimation directe de l'incertitude.

Bibliographie

G. Hoarau et al. (2020), Impact de la masse de particules sur le comportement d'un moniteur de mesure de la contamination atmosphérique (CAM). Congrès Français sur les Aérosols, ASFERA, Paris, France.

Kim, H. & Lee, J. S. (2024). Scalable Monotonic Neural Networks. In The Twelfth International Conference on Learning Representations.

Gawlikowski, J., Tassi, C.R.N., Ali, M. et al. (2023). A survey of uncertainty in deep neural networks. Artif Intell Rev 56 (Suppl 1), 1513–1589.

Clustering methods to segregate avionic failures: A case study from the aerospace industry

Badih Ghattas [1] & Guillaume Poncelin [2] & Alvaro Sanchez San Benito [3]

[1] *Aix-Marseille Université, CNRS, AMSE, France. badih.ghattas@univ.amu.fr*
[2] *Airbus Helicopters, France. guillaume.poncelin@airbus.com*
[3] *I2M, Aix-Marseille Université, Airbus Helicopters, France. alvaro.sanchez.san.benito@airbus.com*

Résumé. Dans cette étude, des méthodes de partitionnement adaptées aux données mixtes ont été employées pour segmenter les codes de panne (CP) avioniques, dans le but de minimiser le délai de remise en état opérationnel des aéronefs. Les équipements avioniques disposent d'une fonction de testabilité leur permettant de surveiller leur état de santé en continu et de signaler toute défaillance détectée par l'envoi d'un CP. Chaque CP enregistré peut correspondre à une vraie ou à une fausse défaillance due à une erreur de monitoring. L'objectif de ce travail est d'identifier les fausses défaillances car elles ont un impact majeur sur le dépannage des aéronefs. Trente jeux de données issus de vols réels, chacun associé à un CP spécifique, ont été manuellement étiquetés par des opérateurs de maintenance afin de mesurer la performance des méthodes comparées : Hierarchical Clustering, convex k-means, Partitioning Around Medoids, k-prototypes, KAMILA et Latent Class Models. HC et convex k-means se sont avérés être les méthodes les plus performantes en atteignant un partitionnement parfait dans 57% des cas.

Mots-clés. Partitionnement, aérospatial, maintenance.

Abstract. In this study, clustering methods for mixed data were used to segregate avionic fault codes (FC) to increase operational availability of the aircraft. Avionic systems are integrated with a testability function that continuously monitors the health status of the aircraft and reports any detected faults through a FC. Each FC may correspond to a real failure or a "false" failure caused by a monitoring error. The objective of this work is to identify false faults, as they have a major impact on aircraft troubleshooting. Thirty datasets from real flights, each associated with a specific CP, were manually labeled by maintenance operators to evaluate the performance of the compared methods: Hierarchical Clustering, convex k-means, Partitioning Around Medoids, k-prototypes, KAMILA, and Latent Class Models. HC and convex k-means emerged as the top performing methods by achieving perfect clustering in 57% of cases.

Keywords. Clustering, aerospace, maintenance.

1 Introduction

Maintenance activities are critical for helicopter operational availability and account for up to one-third of total life cycle costs, making efficient maintenance a key competitive factor in the helicopter industry. The Aircraft Troubleshooting Function, integrated onboard and in ground stations, provides maintenance operators with all the information

required to isolate and remove any detected failure[1]. However, on approximately 10% of all possible fault codes (FC), the detection mean is too sharp and trigger frequent FC detection. The diagnosis step is processed by algorithms built from theoretical knowledge of failure effects. However, this theoretical foundation is often incorrect or incomplete for new aircrafts, resulting in inaccurate procedures that reduce function performance.

To address this challenge, the algorithms must evolve continuously by incorporating real-world failure data recorded after the aircraft entry into service. Although only a small fraction of the FCs recorded are related to a failure, most are a consequence of a monitoring error. Unlike genuine FCs, false FCs demonstrate a consistent and repetitive pattern. Identifying them is a complex, time-consuming task due to the large volume of recorded data and aircraft types.

Clustering analysis offers a promising solution by automatically segregating genuine FCs from false ones. The data includes both continuous and discrete attributes. There are two main approaches for clustering mixed-type data: transforming attributes into a single type for traditional clustering methods like k-means or k-modes, and using clustering methods specifically designed for mixed-type data. The latter approach avoids the pitfalls of arbitrary transformations and potential information loss. Based on benchmark studies (Ghattas 2025), the most effective clustering methods designed for mixed-type data were evaluated on recorded flight data to identify optimal algorithms to support maintenance engineers and improve aircraft availability.

2 Data

Since the recording frequency of FCs varies significantly, data from 500 flights was collected to ensure sufficient coverage. We selected 30 representative FCs, with each corresponding to a separate dataset. The discrete variables are *Phase Appearance* and *Phase Removal*, which indicates the stage of the flight when the FC was recorded and disappear, respectively (ranging from 1 to 5). The continuous variables are *Time Since Phase Appearance (TSPA)* (*Time Since Phase Removal (TSPR)*), which measures the seconds elapsed between the beginning of the flight phase and the recording (disappearance) of the FC; and *Duration*, which represents the total display duration of the FC.

To evaluate and compare clustering performance, maintenance engineers were tasked with manually labeling each FC occurrence in the dataset into two classes. Two scenarios were considered: one with a single class of genuine FC occurrences and a single class of false FC occurrences, and the other with two distinct classes of false FC occurrences. In the absence of ground truth, maintenance engineers labeled the data based on theoretical knowledge and empirical evidence. From this process, the appearance of FC, specifically Phase appearance and TSPA, generally dictates the clustering.

To facilitate a more interpretable distance metric, we rescaled all continuous variables to have a mean of zero and a variance of one prior to conducting the clustering analysis.

[1]Guidelines and methods for conducting the safety assessment process on civil Airborne systems and equipment ARP4761

FC	Mean(standard deviation)			Observed frequency											
	TSPA	TSPR	Duration	Phase appearance					Phase removal					Class	
				1	2	3	4	5	1	2	3	4	5	1	2
1	12.77(30.74)	12.36(31.76)	499.50(478.01)	0	0	3	19	0	0	0	3	0	19	3	19
2	7.93(10.40)	8.46(10.39)	0.56(0.59)	0	335	78	0	0	0	334	79	0	0	78	335
3	7.95(10.36)	8.49(10.32)	0.57(0.59)	0	350	95	0	0	0	349	96	0	0	95	350
4	7.69(9.83)	8.21(9.83)	0.52(0.57)	0	370	91	0	0	0	370	91	0	0	91	370
5	9.34(8.74)	10.80(8.15)	2.34(5.05)	0	1980	1178	0	0	0	1797	1361	0	0	1980	1178
6	9.26(8.45)	10.75(7.80)	2.31(4.97)	0	1947	1204	0	0	0	1775	1376	0	0	1947	1204
7	8.73(8.63)	10.50(7.93)	2.95(5.50)	0	1500	964	0	0	0	1309	1155	0	0	1500	964
8	9.40(8.26)	10.98(7.51)	2.46(5.06)	0	1963	954	0	0	0	1793	1124	0	0	1963	954
9	4.38(7.79)	6.06(7.85)	58.21(44.53)	576	12	0	0	0	36	552	0	0	0	576	12
10	4.57(7.82)	6.05(7.71)	56.81(44.66)	580	21	0	0	0	39	562	0	0	0	580	21
11	2.63(6.29)	41.25(17.71)	115.26(48.50)	537	0	5	0	0	0	0	541	1	0	537	5
12	3.54(6.27)	67.51(31.09)	153.60(57.06)	540	0	5	0	0	3	0	459	81	2	540	5
13	91.64(57.00)	2.87(16.21)	28.19(68.69)	0	0	476	37	0	0	0	6	506	1	37	476
14	16.16(28.33)	13.52(3.63)	18.16(27.08)	649	16	0	0	0	545	120	0	0	0	649	16
15	24.58(5.81)	27.06(11.97)	5.92(5.47)	49	3	0	0	0	43	9	0	0	0	49	3
16	24.63(5.71)	26.81(12.19)	6.07(5.43)	51	3	0	0	0	44	10	0	0	0	51	3
17	24.60(5.76)	27.15(11.89)	5.92(5.46)	50	3	0	0	0	44	9	0	0	0	50	3
18	24.65(5.66)	26.89(12.10)	6.05(5.41)	52	3	0	0	0	45	10	0	0	0	52	3
19	3.91(7.19)	41.67(17.81)	113.80(48.93)	537	0	5	0	4	0	0	541	1	4	537	9
20	3.78(6.33)	67.87(31.08)	153.72(57.11)	540	0	5	0	0	3	0	459	81	2	540	5
21	91.82(57.06)	70.09(54.07)	127.36(149.47)	5	0	17	0	0	2	0	8	12	0	5	17
22	10.89(13.29)	15.62(13.20)	4.73(0.84)	0	254	290	0	0	0	254	290	0	0	254	290
23	17.88(29.97)	18.50(29.92)	0.62(0.61)	0	3	0	0	29	0	3	0	0	29	3	29
24	18.13(30.43)	18.77(30.38)	0.65(0.61)	0	3	0	0	28	0	3	0	0	28	3	28
25	52.57(24.55)	53.96(24.62)	1.39(0.86)	0	0	1	0	190	0	0	1	0	190	1	190
26	71.60(58.97)	72.28(59.01)	0.68(0.57)	0	0	37	0	3	0	0	37	0	3	37	3
27	68.94(59.67)	69.64(59.53)	0.70(0.47)	0	0	31	0	2	0	0	31	0	2	31	2
27	68.94(59.67)	69.64(59.53)	0.70(0.47)	0	0	31	0	2	0	0	31	0	2	31	2
28	49.29(91.94)	50.47(95.09)	1.18(3.20)	0	0	32	1	1	0	0	32	1	1	32	2
29	91.40(56.16)	3.96(23.00)	25.97(37.09)	7	0	485	31	0	7	0	7	509	0	38	485
30	94.68(50.86)	99.95(52.06)	5.26(5.47)	8	0	45	0	20	8	0	45	0	20	8	65

Table 1: Summary statistics of the 30 datasets used in the study.

3 Clustering methods for mixed-type data

Clustering methods for mixed-type data differ in three key aspects: (1) the choice of distance for continuous and categorical variables in distance-based methods, as well as the probability distributions used in probabilistic models, (2) the approach for combining continuous and categorical components, and (3) the technique selected to obtain optimal clusters. Here below, a brief overview of the mechanisms behind the selected methods:

Distance-based methods relies solely on explicit distances or dissimilarities between individuals. With the exception of Hierarchical Clustering, all the approaches discussed in this section follow the same principle as k-means: initializing cluster centers, alternating between a partitioning (assignment) step and a center update step. They differ mainly in the design of the distance used to assign an observation to a cluster.

Notation: K is the number of clusters, n is the population size, p is the total number of continuous variables and q is the total number of categorcal variables.

PAM and HC take a dissimilarity matrix as input. A popular choice is to calculate pairwise Gower's dissimilarities (Gower 1971) among observations. For vectors $\mathbf{x}$ and $\mathbf{y}$:

$$d_{Gower}(\mathbf{x},\mathbf{y}) = \frac{\sum_{j=1}^{p} w_j \left(\frac{|x_j - y_j|}{range(j)}\right) + \sum_{j=p+1}^{p+q} w_j I\{x_j \neq y_j\}}{\sum_{j=1}^{p+q} w_j} \tag{1}$$

where $I\{.\}$ is the indicator function, and $range(j)$ is the absolute difference between values of the j-th variable. If no user specification, the weight w_j is set to 1.

Partitioning around medoids (PAM) (Kaufman 1990) takes the dissimilarity matrix as input to build a partition by affecting observations to the closest "medoid". The K "medoids" are the best representative observation within each cluster in the sense that they will minimise the average dissimilarity with all other points in the cluster.

Ascendant Hierarchical Clustering (HC) (Ward 1963) starts with each observation as its own cluster and iteratively merges the two closest clusters at each step until a single cluster remains. The hierarchy of partitions is represented in a tree structure. At each iteration the distance between the clusters to be merged C_i, C_j and any other cluster C_g is computed to minimise the increase in intra-cluster variance. The aggregation method used in this study to update the dissimilarity matrix is the Ward's algorithm:

$$d(C_i \cup C_j, C_g) = \frac{n_i + n_g}{n_i + n_j + n_g} d(C_i, C_g) + \frac{n_j + n_g}{n_i + n_j + n_g} d(C_j, C_g) + \frac{n_i + n_j}{n_i + n_j + n_g} d(C_i, C_j) \tag{2}$$

where n_i, n_j and n_g are the size of C_i, C_j and C_g respectively.

K-prototypes (Huang 1997) is an iterative clustering method that combines k-means and k-modes. A prototype acts as the equivalent of a medoid in PAM. It uses an hybrid distance function combining squared Euclidean distance for continuous variables and an indicator function for categorical variables. The k-prototypes method minimizes the Total Sum of Distances between observations and the prototype of the cluster (μ_k) to which they belong:

$$TSD(\mu) = \sum_{k=1}^{K} \sum_{i=1}^{n} \mathbf{1}_{ik} \left(\sum_{j=1}^{p} (x_{ij} - \mu_{kj})^2 + \gamma \sum_{j=p+1}^{p+q} I\{x_{ij} \neq \mu_{kj}\} \right) \tag{3}$$

where $\mathbf{1}_{ik} = 1$ if i is in cluster k and 0 otherwise, and γ is the user-specified weight.

Convex k-means (Modha 2003) is an extension of k-means to cluster mixed-type data within a similar framework to k-prototypes, i.e. estimating a weight γ to determine the relative contribution of the categorical attributes. Very similar to (3), it is a convex algorithm as both distance functions used, the squared Euclidean distance and the cosine dissimilarity, are convex. To select the optimal weighting γ^*, convex k-means conducts repeated clustering across a grid of values and identifies the γ that minimises $\mathcal{Q}$, so

$$\mathcal{Q}(\gamma) = \frac{\mathcal{W}_{con}(\gamma)}{\mathcal{B}_{con}(\gamma)} \times \frac{\mathcal{W}_{cat}(\gamma)}{\mathcal{B}_{cat}(\gamma)}. \tag{4}$$

where $\mathcal{W}_{con(cat)}$ is the average within-cluster dispersion of continuous (categorical) attributes and $\mathcal{B}_{con(cat)}$ is the average between cluster dispersion.

Probabilistic models are based on distributional assumptions for the data.

Latent Class Model (LCM) (Everitt 1988) is a mixture model that can also identify important variables for the clustering as well as the optimal number of clusters. Mixture models assume that continuous variables follow a multivariate normal distribution whereas categorical variables follow a multivariate multinomial distribution. Furthermore, the model assumes that variables are independent within clusters. Let $\mathbf{\Omega}$ be the set of relevant variables. For a mixture model with K components, the pdf for an observation $\mathbf{x_i}$ is defined as:

$$f(\mathbf{x_i}|\theta) = \prod_{j \in \mathbf{\Omega}^c} h_j(x_{ij}|\theta_{1j}) \sum_{k=1}^{K} \alpha_k \prod_{j \in \mathbf{\Omega}} h_j(x_{ij}|\theta_{kj}) \tag{5}$$

where $h_j(x_{ij}|\theta_{1j})$ is the distribution function for an irrelevant variable and $h_j(x_{ij}|\theta_{kj})$ is the distribution function for a cluster k with parameters θ_{kj}. Also, $\alpha_k \in [0,1]$ are the mixing coefficients, satisfying $\sum_{k=1}^{K} \alpha_k = 1$. Like for classical mixture models, the EM algorithm is used to find the optimal partition.

KAMILA (Foss 2016) combines k-means and GMM. In fact, the assumption of normality of continuous variables is relaxed by considering a more general spherical distribution with center at the origin. Categorical attributes are assumed to be conditionally independent given cluster membership. Each iteration in KAMILA consists of a partitioning step and an estimation step. KAMILA uses a loglikelihood function as the objective function:

$$H_i(k) = \log\left[\hat{f}_{\mathbf{V}}(d_{ik})\right] + \log \mathrm{P}\left[\mathbf{W}_{ik} = \mathbf{w}_{ik}\right], \tag{6}$$

where $\hat{f}_{\mathbf{V}}$ is a classical kernel estimator, $\mathbf{W}_{ik}$ is the restriction of the discrete variables to k, d_{ik} is the Euclidean distance between i and the cluster center μ_k and P is the multinomial probability function.

In order to measure clustering performance, we use the Adjusted Rand Index (ARI) and the Adjusted Mutual Information (AMI) as ARI is sensitive to highly imbalance cluster size, which is a common scenario (Table 1). All the methods used in the study are available in `R` software (version 4.0.1; `R` Core Team (2020)).

4 Results and discussion

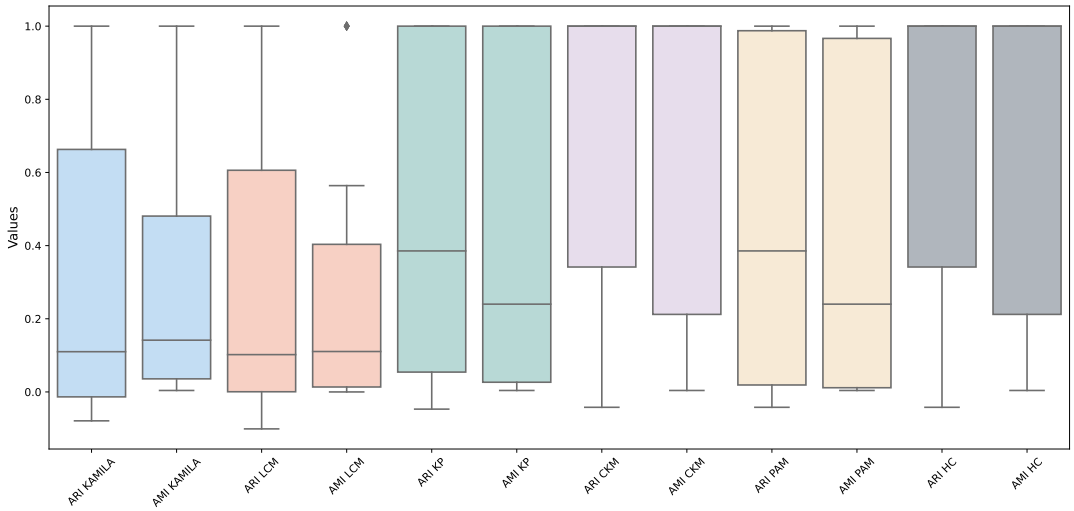


Figure 1: Boxplot of the ARI and AMI results across all datasets for each method.

The results presented in Figure 1 reveal three distinct performance groups. The top-performing group includes convex k-means and HC, both achieving perfect clustering in 57% of the cases. Both methods also attained mean ARI/AMI scores of 0.68/0.65. Furthermore, they demonstrated strong clustering performance across a wide range of datasets, including those with small sample sizes and highly imbalanced clusters, where one cluster contained less than 10% of the observations. The middle-performance group includes k-prototypes and PAM, that achieved a perfect clustering in, respectively, 30%

and 27% of the cases. The mean ARI/AMI scores obtained by k-prototypes are 0.47/0.42 and 0.48/0.45 for PAM. Both methods performed well in the same type of datasets, with not highly unbalanced clusters and sufficiently separated clusters for the continuous variables. Finally, the bottom-performing group includes KAMILA and LCM, both achieving a perfect clustering in 13% and 20% of the cases respectively. Also, the mean ARI/AMI values are 0.31/0.31 for KAMILA and 0.30/0.29 for LCM. Surprisingly LCM is the only method to achieve a perfect clustering for datasets 15,16,17 and 18.

As expected, distance-based methods demonstrate superior overall performance. Despite the diversity of datasets, three specific scenarios are particularly common among avionic fault datasets: highly imbalanced clusters, small sample sizes and high degree of cluster overlap of the continuous variables. Additionally, the continuous variables do not follow a Gaussian distribution. Under these conditions, and consistent with findings from previous benchmarking studies, distance-based methods are better suited for clustering avionic FC records, as probabilistic models depend on distributional assumptions that are not satisfied. Furthermore, the top-performing methods minimise either the intra-cluster variance or the ratio of intra-cluster to between-cluster variance.

Finally, for 80% of the datasets, at least one method achieved perfect clustering. This aligns with empirical evidence suggesting that optimising fault appearance generally results in optimal clustering, as assessed by maintenance engineers.

5 Conclusion

The study provides support to maintenance engineers in two key aspects: (1) offering new insights into the importance of FC appearance, and (2) identifying two effective methods (HC and convex k-means) to assist in segregating genuine FC occurrences from false ones. Additionally, as emphasized in prior research, clustering analysis is highly context-dependent. Consequently, top-performing methods from other studies, such as KAMILA and LCM, should be considered in future analyses, given the wide variety of problems and datasets within the maintenance field of the aerospace industry.

References

Ghattas,B and Sanchez San Benito, A(2025)*Clustering approaches for mixed-type data: A comparative study.*To appear in Journal of Probability and Statistics.

Gower,JC(1971)*A general coefficient of similarity and some of its properties.*Biometrics 27:857–871

Kaufman,L; Rousseeuw,PJ(1990)*Finding Groups in Data: An Introduction to Cluster Analysis.*JohnWiley & Sons,chap 2:68–125

Huang,Z(1997)*Clustering large data sets with mixed numeric and categorical values.*In:Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD),Citeseer,21–34

Modha,DS; Spangler,WS(2003)*Feature weighting in k-means clustering.*Mach Learn 52(3):217–237

Foss,A; Markatou,M; Ray,B and Heching, A.(2016) *A semiparametric method for clustering mixed data.*Mach Learn 105(3):419–458

Ward,J.H.(1963)*Hierarchical grouping to optimize an objective function.*J.Am.Stat.Assoc.58(301),236–244.

Everitt,B.S.(1988)*A finite mixture model for the clustering of mixed-mode data.*Stat.Prob.Lett.6(5),305–309.

Identification of Graph Features to Improve GNN Performance

Itai Zehavi[*,+1] Eloi Campagne[*,+] Argyris Kalogeratos[*]

[*] *Centre Borelli, Université Paris-Saclay, CNRS, Ecole Normale Supérieure Paris-Saclay*
[+] *EDF R&D, Palaiseau*

Résumé. Les Graph Neural Networks (GNNs) sont de plus en plus utilisés pour modéliser les séries temporelles spatialisées, notamment pour la prédiction de la consommation électrique. La qualité du graphe sous-jacent est cruciale pour leurs performances Campagne et al. (2024), mais son évaluation est coûteuse en ressources. Nous proposons une méthode pour analyser la pertinence des graphes en amont, en utilisant des caractéristiques structurelles comme la régularité Shuman et al. (2013) et la bande passante Daković et al. (2019). En comparant ces caractéristiques entre graphes réels et aléatoires, nous identifions des structures informatives. Nous appliquons également ces caractéristiques sur des séries temporelles compressées pour capturer des dynamiques temporelles complexes. Nos méthodes offrent des informations complémentaires et non redondantes. Toutefois, il reste à déterminer quelles caractéristiques améliorent effectivement les performances des GNNs.

Mots-clés. Théorie des Graphes, Séries Temporelles, Modélisation Statistique, Réseaux de Neurones Graphiques

Abstract. Graph Neural Networks (GNNs) are increasingly used to model spatio-temporal series, for predicting electricity consumption among other applications. The quality of the underlying graph is crucial for their performance, but evaluating it is resource-intensive. We present a method to analyze the relevance of graphs upstream, using structural features such as the smoothness and bandwidth of the time series over a given graph. By comparing these features between real and random graphs, we identify informative structures. We also apply these features to compressed time series to capture complex temporal dynamics. Our approach provides complementary and non-redundant information, however, it remains to be determined which features actually enhance GNN performance.

Keywords. Graph Signals, Time Series, Statistical Modeling, Graph Neural Networks

1 Introduction

GNNs have emerged as a powerful tool for modeling structured data, particularly in applications such as spatio-temporal forecasting. The performance of GNNs is highly dependent on the structure of the underlying graph. Poorly constructed graphs may introduce noise or irrelevant connections and lead to suboptimal predictions (Campagne et al., 2024).

To address this issue, we investigate the quality of a given graph using structural features such as the *smoothness* and *bandwidth* of the graph signals induced by the time series. These

[1]Mail: itai.zehavi@ens-paris-saclay.fr

features allow us to assess how well a graph represents meaningful relationships between its nodes before it is used for training a GNN. By comparing the distribution of these features in real graphs to those generated randomly, we identify whether a given graph possesses informative structures that can enhance GNN performance. Additionally, we extend our study to compressed time series, aiming to capture richer temporal dependencies.

This paper formulates the problem of graph structure evaluation in the context of GNNs, defines key metrics, and outlines our methodology for feature extraction and comparison.

2 Preliminaries

Definition 2.1 (Smoothness, see Shuman et al. (2013))**.** The smoothness of a graph signal $\mathbf{x} \in \mathbb{R}^N$ defined on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with weight matrix $\mathbf{W}$ is quantified using the quadratic form of the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$:

$$S(\mathbf{x}) = \mathbf{x}^\top \mathbf{L}\mathbf{x}, \tag{1}$$

where $\mathbf{D}$ is the diagonal degree matrix. A lower value of $S(\mathbf{x})$ implies that adjacent nodes have similar values, indicating high smoothness, while a higher value signifies strong variations across edges.

Definition 2.2 (Bandwidth, see Daković et al. (2019))**.** The bandwidth of the local smoothness for a vertex-frequency distribution $G(n, k)$ is defined as:

$$\text{Bandwidth} = \frac{1}{|\mathcal{V}|} \sum_{n \in \mathcal{V}} \sqrt{\frac{\sum_{k=1}^{N} (\lambda_k - \lambda(n))^2 G(n,k)}{\sum_{k=1}^{N} G(n,k)}}, \tag{2}$$

This measures the average dispersion of $G(n, k)$ around $\lambda(n)$ across all nodes, indicating whether the energy is concentrated in a narrow range or spread across multiple frequencies. $\lambda(n)$ corresponds to the average frequency of the signal at node n. The vertex-frequency distribution $G(n, k)$ represents the energy of the signal x at node n and frequency λ_k.

3 Results

Given a time series dataset with observations distributed at different nodes, a graph can be constructed where edges represent relationships between those nodes (here each node is a region of France). The goal is to determine whether the structure of this graph influences GNN performance and to identify which characteristics of the time series contribute to better predictive capabilities.

We perform a comparative analysis by generating random graphs and computing the distributions of chosen measures such as those in Section 2. A significant deviation between the distributions of real and random graphs suggests the presence of informative structures, as illustrated in Figure 1.

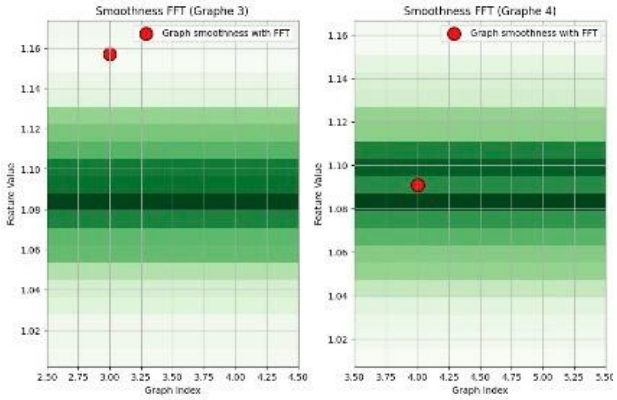


Figure 1: Monte Carlo simulation of the smoothness distribution. The background green shades represent the probability distributions, while the red dot indicates the feature value of a given graph. Left: a real graph with an informative structure related to the observed feature. Right: a less informative graph appears to behave no different than randomly generated graph structures for the considered time series data.

Since each node in our dataset is associated with a time series, applying features directly to the raw signals only captures relationships between events occurring at the same instant. To gain a richer understanding of our data, we compress the signals and compute features that incorporate information from multiple time instants. This approach allows us to analyze temporal dependencies beyond immediate correlations. Furthermore, we observe that the extracted features are decorrelated, meaning that they provide complementary information about the graph structure. This confirms that different features capture distinct properties of the graph.

In conclusion, our work establishes a methodology for identifying features for which the graph is a statistically informative structure. Future research could focus on correlating the values of these features with GNN performance to determine which features best indicate graph quality. Ultimately, this could lead to the construction of graphs tailored to the data for improved predictive accuracy.

References

Campagne, E., Amara-Ouali, Y., Goude, Y., and Kalogeratos, A. (2024). Leveraging Graph Neural Networks to Forecast Electricity Consumption. *European Conference on Machine Learning and Knowledge Discovery in Databases, Vilnius. Machine Learning for Sustainable Power Systems Workshop.*

Daković, M., Stanković, L., and Sejdić, E. (2019). Local Smoothness of Graph Signals. *Mathematical Problems in Engineering.*

Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., and Vandergheynst, P. (2013). The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains.

Statistique pour la société

Vers une carte des élections basées sur des évaluations

Jean-Baptiste Aubin [1] & Irène Gannaz [2] & Antoine Rolland [3]

[1] *DEEP, INSA-Lyon, ICJ, France, jean-baptiste.aubin@insa-lyon.fr*
[2] *Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP, 38000 Grenoble, France, irene.gannaz@grenoble-inp.fr*
[2] *ERIC EA 3083, Université de Lyon, Université Lumière Lyon 2, France, antoine.rolland@univ-lyon2.fr*

Résumé. Cet article explore la modélisation des élections basées sur des évaluations. Il généralise les indices caractérisant les modélisations d'élections par ordres de préférences introduits dans la littérature aux élections par évaluations. Il introduit trois indices (unanimité, indifférence, dimension). Il ouvre la voie vers une une mesure de dissimilarité entre élections et une cartographie de celles-ci.

Mots-clés. Élections par évaluation, dissimilarité entre élections, indices de structure électorale, analyse en composantes principales.

Abstract. This paper explores the modeling of elections based on evaluations. It extends extending indexes which caracterize preferences based election modeling to evaluations based elections. It introduces three indices (unanimity, indifference, dimension) and a dissimilarity measure between elections. The approach paves the way to better map the election space.

Keywords. Evaluation-based elections, election dissimilarity, electoral structure indices, principal component analysis.

Introduction

Notre étude a pour cadre la théorie du choix social computationnel. Cette discipline se concentre sur l'analyse de la prise de décision collective, en particulier sur les aspects computationnels de l'étude des élections.

Formellement, une élection est définie par un ensemble de candidats et un ensemble d'électeurs ayant certaines préférences sur ces candidats. Dans le cadre classique de la théorie du choix social, les électeurs sont capables de fournir une relation de préférence sur les candidats (typiquement un ordre total).

La théorie des élections rencontre de nombreux défis, tels que la détermination du ou des gagnants d'une élection, le calcul des marges de victoire, l'analyse des différentes formes de manipulation, l'étude des propriétés axiomatiques (par exemple, le critère du vainqueur de Condorcet), et bien d'autres encore.

Afin de relever ces défis, il est important de pouvoir caractériser les élections. On définit une élection comme un ensemble de préférences, ou d'évaluations, de votants sur des candidats. Identifier les formes possibles d'élections et leurs principales caractéristiques est utile pour comprendre les différences de comportement des modes de scrutin. Dans cet objectif, Szufa *et al.* (2020) ont introduit un cadre méthodologique, appelé carte des élections, qui permet une représentation graphique facilitant la compréhension de l'espace des élections basées sur des ordres de préférence. A partir d'un ensemble d'élections donné, les distances entre chaque paire d'élections sont calculées, puis les élections sont projetées dans un espace euclidien bidimensionnel de manière à ce que les distances euclidiennes reflètent le plus fidèlement possible celles obtenues avec la métrique grâce à une projection (*"multidimensional scaling"*).

Dans notre recherche, nous nous concentrons sur le cadre alternatif où les préférences des électeurs sur les candidats sont exprimées sous forme d'évaluations : les électeurs donnent une note à chacune des alternatives proposées. Les distributions de ces préférences peuvent être modélisées de différentes manières, comme détaillé par exemple dans Rolland *et al.* (2024). Le but de ce travail est de généraliser les cartes d'élections précédemment citées au contexte des élections basées sur les évaluations.

1 Indices décrivant des élections par évaluations

Boehmer *et al.* (2021) ont introduit des "élections type" afin d'aider à l'interprétation du positionnement d'élections dans cette carte pour des élections basées sur des ordres de préférence. Nous les généralisons ici au cas des élections basées sur des évaluations. Supposons que les évaluations de n votants sur m candidats soient résumées dans une matrice $m \times n$ nommée Φ. Dans la suite, on confondra élection et matrice des évaluations associée, Φ.

Unanimité. Une élection Φ basée sur des évaluations est dite unanime (entre les votants), si chaque candidat reçoit toujours la même évaluation :

$$\exists a_1, \ldots, a_m \in [0,1] \text{ tel que } \forall i = 1, \ldots, m, \forall j = 1, \ldots, n, \Phi(i,j) = a_i.$$

Indifférence. Une élection Φ est dite "indifférente" si tous les votants sont indifférents par rapport à tous les candidats :

$$\exists a_1, \ldots, a_n \in [0,1] \text{ tel que } \forall i = 1, \ldots, m, \forall j = 1, \ldots, n, \Phi(i,j) = a_j.$$

Polarisation. Une élection est dite "polarisée" si exactement la moitié des votants ont un profil d'évaluations $(\Phi(1,j))_{j=1,\ldots,n}$ et l'autre moitié a le profil opposé $(\Delta - \Phi(1,j))_{j=1,\ldots,n}$ où Δ désigne la note maximale pouvant être donnée (dans la suite, Δ sera pris égal à 1 sans perte de généralité).

Remarque : Dans le cas extrême(ment central) où $\forall (i,j) : \Phi(i,j) = 1/2$, l'élection est à la fois polarisée, indifférente et unanime.

Nous considérerons par la suite également le nuage de points associé à une élection Φ, qui permet de mieux visualiser les notions. Chaque électeur est représenté par un point

en dimension m, dont la i-ème coordonnée est donnée par le vote pour le i-ème candidat. L'élection est donc un nuage de n points en dimension m (ou de m points en dimension n).

1.1 Indice d'unanimité

Définition d'un indice d'unanimité
L'indice d'unanimité U d'une élection Φ est défini comme

$$U(\Phi) := 1 - \frac{4}{m}\sum_{i=1}^{m} \mathrm{var}(\Phi(i,.)).$$

On a le résultat suivant :

Propriété 1 *Pour toute élection Φ dont les évaluations données varient entre 0 et 1,*

$$0 \leq U(\Phi) \leq 1.$$

Et $U(\Phi) = 1$ si et seulement si Φ est une élection unanime ($\exists a_1, \ldots, a_m \in [0,1]$ tel que $\forall i = 1, \ldots, m, \forall j = 1, \ldots, n, \Phi(i,j) = a_i$).

Cet indice est inversement proportionnel à la variance totale du nuage des n points associé à Φ, et, dans le cas d'*unanimité*, les n points sont confondus : tous les électeurs donnent les mêmes notes. L'indice vaut 0 si toutes les variances candidat par candidat sont maximales, c'est-à-dire valent $\frac{1}{4}$, ce qui correspond au cas où, pour chaque candidat, la moitié des votants donne 1 et l'autre moitié, 0.

1.2 Indice d'indifférence

Définition d'un indice d'indifférence
L'indice d'indifférence I d'une élection Φ est défini comme

$$I(\Phi) := 1 - \frac{4}{n}\sum_{j=1}^{n} \mathrm{var}(\Phi(.,j)).$$

On a le résultat suivant :

Propriété 2 *Pour toute élection Φ dont les évaluations données varient entre 0 et 1,*

$$0 \leq I(\Phi) \leq 1.$$

De plus, $I(\Phi) = 1$ si et seulement si Φ est une élection indifférente ($\exists a_1, \ldots, a_n \in [0,1]$ tel que $\forall i = 1, \ldots, m, \forall j = 1, \ldots, n, \Phi(i,j) = a_j$.).

Cet indice vaut 1 si les distributions des votes de chaque électeur est constante. Chaque votant donne la même note à tous les candidats. Par ailleurs, cet indice vaut 0 si toutes les variances électeur par électeur valent $\frac{1}{4}$, ce qui correspond au cas d'extrême désaccord où, pour chaque électeur, la moitié des candidats a 1 et l'autre moitié, 0.

1.3 Indice de dimension

Soient $V_1, \dots V_m$ les valeurs propres associées à la matrice de variance-covariance de la transposée de Φ. En d'autre terme, V_1 est la variance portée par le premier axe d'une Analyse en Composantes Principales (ACP) sur le tableau votants / candidats. Cette représentation permet de voir si la distribution des évaluations des votants sur les candidats est anisotrope ou si elle possède une dimension privilégiée.

Définition d'un indice de dimension
Soient $V_1, \dots V_m$ les valeurs propres associées à la matrice de variance-covariance de la transposée de Φ. L'indice de dimension D est défini comme

$$D(\Phi) := \frac{m \times \frac{V_1}{\sum_{i=1}^{m} V_i} - 1}{m - 1}$$

Cet indice vaut 1 si le nuage de points associé à une élection Φ a une unique direction. Il peut être vu comme une distance à l'anisotropie.

Propriété 3 *Pour toute élection Φ,*

$$0 \leq D(\Phi) \leq 1.$$

*Pour une élection où les préférences sont unidimensionnelles (*Single peaked*) $D(\Phi) = 1$. En particulier, pour une élection indifférente ou polarisée, $D(\Phi) = 1$. De plus, dès qu'il existe 3 votants qui ont des profils d'évaluations non colinéaires, $D(\Phi) < 1$.*

Ce résultat s'explique par le fait que, lors d'une élection polarisée, les préférences de tous les votants, réduites à deux positions, sont donc sur une même droite, concentrant toute l'information sur la première composante principale.

Pour étendre cette notion, on pourrait essayer de définir une "dimension intrinsèque" d'une élection.

2 Une première carte d'élections

Considérons le cas avec $m = 4$ candidats et $n = 100$ votants. Nous avons généré 22000 élections Φ. Pour chaque élection simulée, nous avons calculé les indices d'unanimité U, d'indifférence I et de dimension D. Les résultats sont données en Figure 1. Les méthodes de simulation utilisées sont décrites dans Rolland *et al.* (2024). Nous avons considéré des distributions $\{\Phi(., j),\ i = 1, \dots, m\}$:

- Fixes, c'est-à-dire telles que chaque votant j donne une même note $\Phi(i, j) = e_j$ à chaque candidat - en rouge sur la Figure 1 ;
- Indépendantes et Identiquement Distribuées (IID) avec des lois uniformes ou des lois Beta - en bleu clair sur la Figure 1 ;

- Dépendantes et Identiquement Distribuées (DID) avec des lois uniformes ou des lois Beta - en bleu foncé sur la Figure 1 ;
- Indépendantes et non identiquement Distribuées (IDD)
  - avec des lois Beta - en violet sur la Figure 1,
  - avec des distributions de *Polarisation Parfaite* - en jaune sur la Figure 1 ;
- Dépendantes et non identiquement Distribuées (DDD) avec des lois Beta - en noir sur la Figure 1.

Les paramètres des lois Beta ont été choisis aléatoirement selon des lois uniforme sur [0,5]. Les lois avec dépendance ont été obtenues en utilisant des copules Gaussiennes, tel que décrit dans Rolland *et al.* (2024). Les matrices de corrélation utilisées dans le copules ont été généres selon des lois uniformes sur l'ensemble des matrices de corrélation, à laide de la librairie `R randcorr`. Notons que les distributions des votants $\{\Phi(.,j),\ j=1,\dots,n\}$ sont indépendantes, sauf dans le cas de distributions fixes.

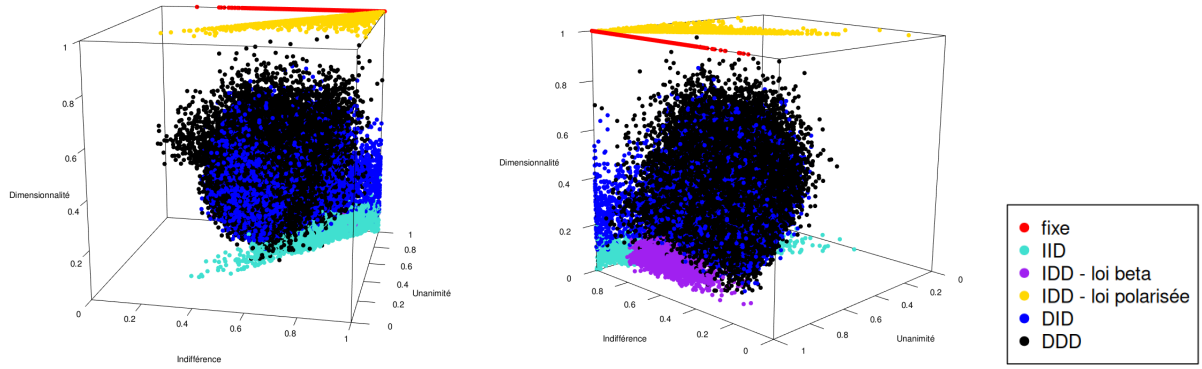


Figure 1: Représentation des élections Φ simulées en fonction des indices d'unanimité U, d'indifférence I et de dimension D. Chaque point représente une élection. Les 2 figures montrent le même nuage de points, avec un angle différent.

On observe que les distributions fixes ont comme attendu un indice de polarité de 1 et un indice d'accord de 1. Les distributions de *Polarisation Parfaite* ont également un indice de polarisation de 1. Inversement, l'indice de polarisation est faible quand les données sont indépendantes.

Perspectives

Ce travail introduit des indicateurs caractérisant les élections. L'objectif est de cartographier les élections par évaluations comme cela a été fait dans Boehmer *et al.* (2021) pour les élections basées sur des ordres de préférences. Des perspectives de ce travail sont donc de construire cette cartographie, en déduisant des caractéristiques des mesures de dissimilarité

entre les élections des indices introduits ici. D'autres mesures de dissimilarité pourraient être exploitées, comme par exemple les fonctions de divergences type Kullback-Leibler (Hellinger, Jensen-Shannon, etc), les distances basées sur le transport optimal (Wasserstein, Sickhorn, etc), ou les statistiques-test associées à des tests d'adéquation (Khi-2, Kolmogorov-Smirnov, Anderson-Darling, etc). Une fois la cartographie établie, nous souhaitons réaliser une étude du comportement des méthodes de vote sur des simulations, un des objectifs étant d'observer les différences entre les modes de scrutin selon le contexte de l'élection.

Bibliographie

Boehmer N., Bredereck R., Faliszewski P., Niedermeier R., and Szufa S. (2021) Putting a compass on the map of elections. In Proceedings of IJCAI-2021,

Faliszewski P., Kaczmarczyk A. & Sornat K. (2023). Diversity, Agreement, and Polarization in Elections. In Proceedings of IJCAI-2023,

Rolland A., Aubin J. B., Gannaz I., & Leoni S. (2024). Probabilistic models of profiles for voting by evaluation. Social Choice and Welfare, 63(2), 377-400,

Szufa S., Faliszewski P., Skowron P., Slinko A., Talmon N. (2020). Drawing a Map of Elections in the Space of Statistical Cultures. In Proceedings of AAMAS-2020.

Unpacking the Map of Opportunity: a multidimensional approach to the sub-national variations of social mobility in France

Marina Gomtsyan[1] & Cyril Jayet [2] & Aurélie Fischer [3]
Floriana Gargiulo [4] & Maxime Lenormand [5]

[1] *CNRS/Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, France et mgomtsian@lpsm.paris*
[2] *CNRS/Groupe d'étude des méthodes de l'analyse sociologique de la Sorbonne (GEMASS), Sorbonne Université et cyril.jayet@sorbonne-universite.fr, louis-andre.vallet@cnrs.fr*
[3] *CNRS/Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Université Paris Cité, France et aurelie.fischer@lpsm.paris*
[4] *CNRS/Groupe d'étude des méthodes de l'analyse sociologique de la Sorbonne (GEMASS), France et floriana.gargiulo@cnrs.fr*
[5] *AgroParisTech/Cirad/CNRS/INRAE/Territoires, environnement, télédétection et information spatiale (TETIS), Université de Montpellier, France et maxime.lenormand@inrae.fr*

Résumé. Les recherches économiques et sociologiques ont montré que la force de la mobilité sociale varie considérablement entre les régions d'un même pays. Cette approche ne tient toutefois pas compte de la nature multidimensionnelle de la mobilité des classes sociales. L'objectif de cette étude est de mettre en lumière l'importance de cette multidimensionnalité pour rendre compte des inégalités géographiques. Nous utilisons les données de l'enquête française sur les forces de travail menée entre 1982 et 2022 (1,276,752 individus). Nous mesurons la classe sociale à l'aide d'une version à six classes de la classification socio-économique européenne et nous construisons 96 tables de mobilité, une pour chaque département de naissance. Nous proposons une approche analytique originale en utilisant des techniques de réduction de la dimensionnalité (analyse en composantes principales et analyse factorielle des correspondances), des méthodes de clustering et des modèles topologiques log-linéaires pour identifier les principales dimensions qui structurent les flux de mobilité entre les classes sociales. Nous identifions tout d'abord des modèles de mobilité spécifiques correspondant à trois grands types de territoires : les territoires ruraux, les territoires urbains à cols blancs supérieurs et les territoires urbains industriels. Cette distinction fournit une carte géographique de la mobilité sociale en France très différente d'une approche unidimensionnelle classique. Deuxièmement, nous montrons que la force de l'héritabilité des différentes classes n'est pas bien corrélée et que certaines sont négativement corrélées. La fluidité ascendante et la fluidité descendante sont également négativement corrélées, et toutes deux sont faiblement corrélées à l'héritabilité. Par conséquent, les territoires diffèrent non seulement par le degré de fluidité sociale, mais aussi par les formes prédominantes d'héritabilité et par les caractéristiques spécifiques de la fluidité sociale.

Mots-clés. Mobilité sociale, clustering, réduction de la dimensionnalité, modèle unidiff

Abstract. Economics and sociological research have shown that the strength of social mobility substantially varies between regions within countries. Such an approach misses however the multidimensional nature of social class mobility. The objective of this paper is to shed light on the importance of this multidimensionality to account for geographical inequalities. We use data from the French labor force survey conducted between 1982 and 2022 (1,276,752 individuals). We measure social class with a six-class version of the European socioeconomic classification and we construct 96 mobility tables, one for each region of birth. We propose an original analytical approach by using dimensionality reduction techniques (principal components analysis and correspondence analysis), clustering methods, and topological log-linear models to identify the main dimensions that structure the flows of mobility between social classes. We firstly identify specific patterns of mobility corresponding to three main types of territory: rural territories, higher white collar urban territories, and industrial urban territories. This distinction provides a very different geographical map of France's social mobility than a classical unidimensional approach. Secondly, we show that the strength of the inheritance of the different classes are not well correlated and that some are negatively correlated. Upward fluidity and downward fluidity are also negatively correlated, and both are weakly correlated to inheritance. Therefore, territories differ not only in the degree of social fluidity but also in the predominant forms of inheritance and in the specific characteristics of social fluidity.

Keywords. Social mobility, clustering, dimensionality reduction, unidiff model

1 Introduction

Social mobility is the movement of individuals from one social class to another. Subnational comparisons of social mobility within a single country have so far relied on unidimensional measures of the strength of mobility. Such an approach misses however the multidimensional nature of social class mobility. For example, social mobility in a territory can be weak because children of managers become managers, but also because children of farmers become farmers. Similarly, a strong mobility in a region can mean very different things: people born in agricultural regions may have different types of mobility than those living in urban or industrial regions. Distinguishing those different types of mobility is fundamental to better understand how life trajectories are constrained by the territory in which someone is born. We shall therefore analyse how the type of mobility differs by regions of birth using an approach that account for the multidimensionality of social mobility.

In this work, we analyse French data from the Labor Force Survey from 1982 to 2022 (1,276,752 individuals). We measure social class using a six-class version of the European socioeconomic classification. For social origin, we use the father's social class. We construct 96 mobility tables, one for each region of birth. Mobility tables represents frequencies in each origin-destination (also referred as position) combination. We employ original methods to analyse the mobility tables of each region, combining dimensionality reduction techniques (correspondence analysis, principal component analysis), cluster analysis, and topological log-

linear models. The region is the French "département" where the person is born. There are 96 "departments" in metropolitan France. Firstly, we identify the main structural divisions in the French territory: the opposition between rural and urban region and the opposition between white collar urban regions and industrial urban regions. Secondly, to evidence the multidimensionality of social fluidity, we investigate the correlation between the inheritances of the different classes, downward fluidity, and upward fluidity.

To measure social fluidity at the regional level, we shall use the log-multiplicative uniform difference model (unidiff), see Erikson and Goldthorpe [1992]. It is defined as follows

$$\log(\mu_{ijk}) = \lambda + \lambda_i^O + \lambda_j^P + \lambda_k^D + \lambda_{ik}^{OD} + \lambda_{jk}^{PD} + \phi_k \psi_{ij},$$

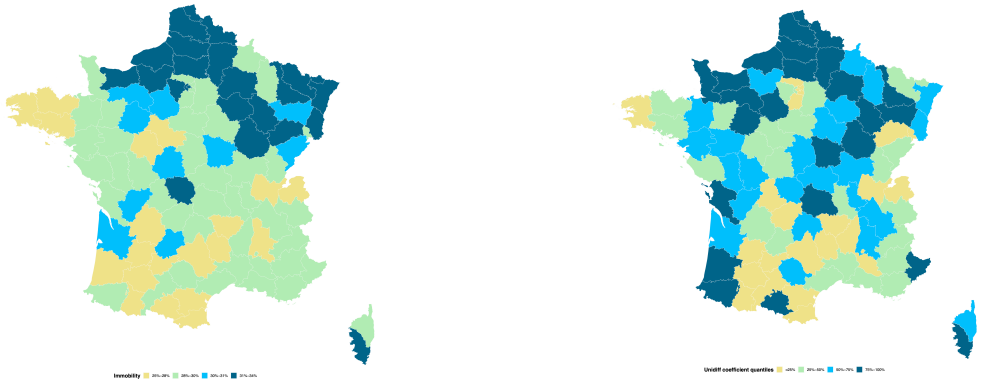
where μ_{ijk} is the expected frequency in the cell ijk, i being the origin, j the position, and k the department, λ is called the main effect, λ_i^O the marginal effect of origin, λ_j^P the marginal effect of position, λ_k^D to the marginal effect of the department. λ_{ik}^{OD} and λ_{jk}^{PD} account for the regional change of the marginal distributions from one department to another. ψ_{ij} is a matrix of coefficients describing the association between origin and position. ϕ_k gives one coefficient for each department (the layer coefficients) to describe the difference of associations between them. Social fluidity will be measured by the layer coefficient associated to each department. Upward and downward fluidity describe the movement in the economic hierarchy that is more than would be expected by the marginal distributions. For more details on measuring upward and downward fluidity we refer the reader to the core fluidity model, see Erikson and Goldthorpe [1992].

2 Results

Figures 1a and 1b present the geographical variations of absolute mobility and of social fluidity. The maps reveal that regions in the north and eastern-north tend to be the most immobile and less fluid ones.

To identify different types of mobility, we performed a correspondence analysis on the table crossing the region of birth and 36 variables measuring the number of individuals in each cell of the mobility table. The correspondence analysis (CA) reduces those 36 variables to a smaller set of factors that explain most of the variations between the mobility tables. Figure 2 presents the geometric result of the CA. Table 1 provides the correlations between the coordinates on the different dimensions obtained by the CA and the socioeconomic characteristics of the regions.

A large part (83.4%) of the regional variations in the mobility tables can be explained by only two orthogonal factors. The first one explains 58.8% of the variance and the second one 24.6%. This first factor can be interpreted as opposing the rural regions to the richest urban regions, while the second factor opposes those two last types of regions to the urban blue-collar regions. At the top right of Figure 2, variables correspond to mobility from or toward the group of farmers. In these regions, the importance of the class of farmers implies that children of farmers have more chances to stay in this class, but also that people from different origins can become farmers.



(a) Regional variations of immobility (b) Regional variations of unidiff coefficient

Figure 1: The colors correspond to the different quartiles: in dark blue, the 25% more immobile or less fluid regions, in yellow the most mobile or fluid ones. Light blue (50-75%) and green (25-50%) are the two intermediate groups

	Dimension 1	Dimension 2
GDP	-0.66	0
Unemployment	-0.24	-0.37
Urbanisation	-0.81	-0.24
Density	-0.46	0.13

Table 1: Correlations between axes coordinates and socioeconomic characteristics of regions.

In the left part of Figure 2 and, therefore, opposed to the farming regions, we find the richest regions of France: Paris, the regions around Paris (Haut de Seine, Yvelines, Val de Marne), as well as the regions containing France's main cities: Haute Garonne (Toulouse), Alpes-Maritimes (Nice), Rhone (Lyon), etc. Those regions are characterised by the importance of the social reproduction of the class of managers and professionals, but also all the flows from and toward this class. The position of the regions on the first axis strongly correlates with its urbanisation (- 0.8), but also its Gross domestic product (GDP) (-0.66), and to a lesser extent to its density (-0.46).

The second factor (from the bottom of the graphics to the top) opposes the farmer regions and the urban richer regions that we already described to the northern and eastern regions, such as Nord, Moselle, and Pas-de-Calais. This time, unemployment is the most correlated socioeconomic variable (-0.37), whereas GDP is not correlated at all.

We have performed a cluster analysis based on the proximity between the regions on the first five factors obtained by the correspondence analysis. Figure 3 represents the map of the regions of France coloured by the clusters we obtained. This last map, compared to Figure 1a, shows how a multidimensional approach to social mobility gives a different picture than a unidimensional one. It reveals that the different types of mobility correspond to different and identifiable geographical areas of France. The first cluster in dark green corresponds to the

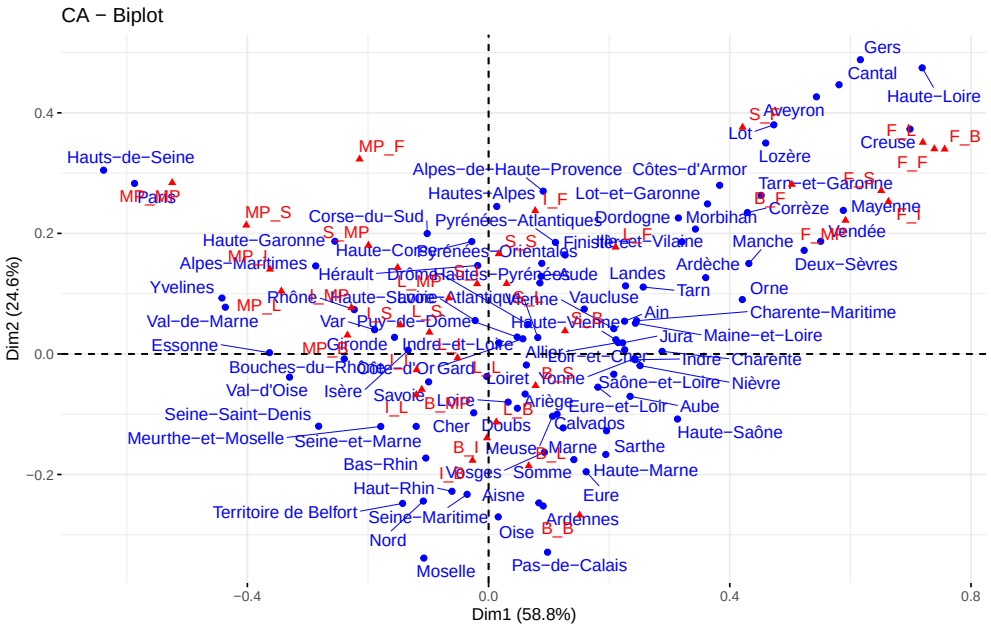


Figure 2: First two dimensions of a correspondence analysis on the 96 mobility tables. The letters in red represent the cells of the mobility table. The letter before the "_" is the origin, after it is the destination. L: Lower White Collar; MP: Manager-Professional; F: Farmers; S: Self-employed; B: Blue collar; I: Intermediary occupations.

regions with greater mobility from and towards the category farmer. This corresponds to the center-south of France and to the west. The second cluster (in light green) corresponds to the richest regions and mobility from and toward the category of Managers and Professionals. The cluster in dark blue corresponds to the industrial regions of the north and east. The last cluster, in light blue, corresponds to more intermediate situations between those three clusters.

To conduct a multidimensional analysis of social fluidity, we first compute for each region a topological fluidity model, see Erikson and Goldthorpe [1992]. We obtain for each region 8 parameters describing the strength of each class inheritance, and of upward and downward fluidity. We examine to what extent the inheritance of each class and the two types of fluidity are correlated to one another. To conduct such an analysis, we have performed a principal component analysis (PCA) on the 8 variables describing the different dimensions of social fluidity. Figure 4 provides the graph of the PCA drawn with the first two components. Variables measuring the economic characteristics of the regions are projected on the graph but they do not contribute to the construction of the components. The graph represents the correlations between the different social mobility tables' characteristics, the angle between two variables corresponding to the correlation (the more acute the angle, the stronger the positive correlation). Table 2 provides the correlations between the position of the region on each component and the socioeconomic characteristics of the regions.

The graph and the table reveal several key insights. Firstly, class inheritance is not a

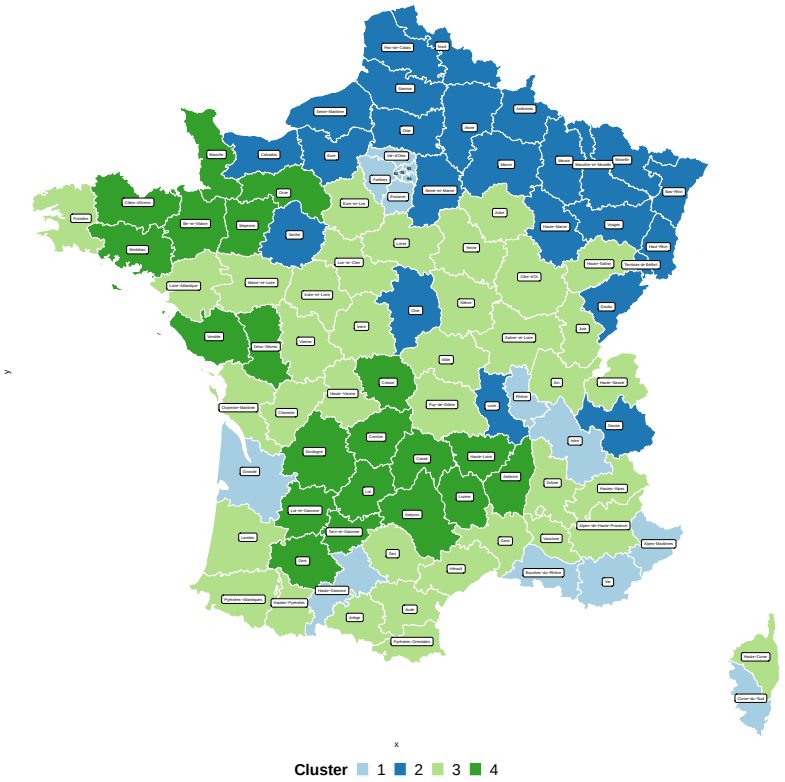


Figure 3: Map of the different clusters of mobility

	Dimension 1	Dimension 2
GDP	-0.65	-0.15
Unemployment	-0.03	0.18
Urbanisation	-0.53	-0.23
Density	-0.58	-0.20

Table 2: Correlation between socioeconomic variables and components' coordinates.

unidimensional phenomenon; otherwise, all the inheritance variables would have been pointing in the same direction. This first component (horizontal axis) synthesizes the positive association between the inheritance of managers and professionals and the inheritance of blue collars. This inheritance is opposed to the inheritance of lower non-manual class and farmers, showing that there is a negative association between these latter variables and the two first ones. The variables measuring the inheritance of the intermediary occupation and of self-employed occupations are orthogonal to the first component, meaning that there is rather weak correlation between them and the other inheritances. Territories differ, therefore, not only in the strength of their social fluidity but in the type of inheritance that prevails within them. Secondly, one might have expected inheritance to conflict with fluidity, and upward and downward fluidity to be positively correlated to form a general level of fluidity.

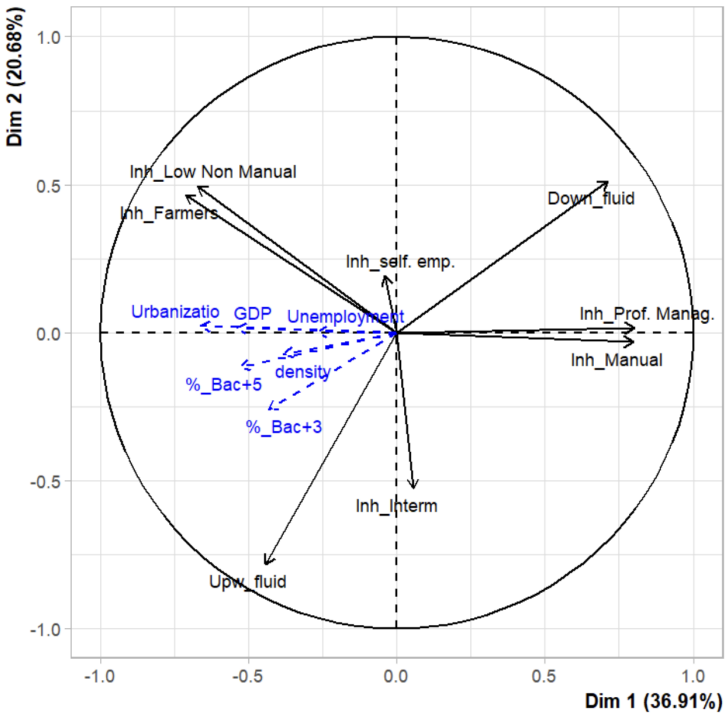


Figure 4: PCA on the dimensions of social fluidity (Components 1 & 2). Arrows represent variables that are parameters obtained in topological log-linear models estimated separately for each region. Variables starting by "Inh" correspond to the inheritance of a class in a region.

The second component (vertical) evidences on the contrary the negative association between upward and downward fluidity. A general level of fluidity hides, therefore, different types of fluidity. Thirdly, the position on the first component is negatively correlated to the GDP, the urbanization, and the density. The axis opposes, therefore, rural areas on the right to richer urban areas on the left. The strength of the inheritance of the position of blue-collar and of manager are therefore correlated to a lower GDP and to a lower urbanization. Those results show that more positions of managers and professionals in richer areas give more opportunity for everyone to attain those positions and not only to the children of managers and professionals. On the contrary, unemployment in blue collar regions is more correlated to the second component and thereby to downward fluidity.

References

R. Erikson and J. H. Goldthorpe. *The constant flux : a study of class mobility in industrial societies*. Oxford University Press, 1992.

Money laundering detection: a contrastive learning approach on financial time series with a transformer

Harold Guéneau [1] & Alain Celisse [2] & Pascal Delange [3]

[1] *Université Paris 1 - Panthéon-Sorbonne & Shine, France, harold.gueneau@shine.fr*
[2] *Université Paris 1 - Panthéon-Sorbonne, France, alain.celisse@univ-paris1.fr*
[3] *Marble, France, pascal@checkmarble.com*

Résumé. Les institutions financières ont l'obligation de lutter contre le blanchiment d'argent, mais les approches traditionnelles basées sur des systèmes à base de règles s'avèrent souvent inefficaces et coûteuses à exploiter en raison d'un compromis difficile entre le taux de détection et les faux positifs. Ce travail propose une nouvelle méthodologie en deux étapes pour la détection du blanchiment d'argent, exploitant des données temporelles avec un transformeur. La première phase se concentre sur le développement de représentations robustes, indépendantes des labels de fraude disponibles, grâce à l'apprentissage contrastif. La deuxième phase utilise ces représentations pour générer un score de blanchiment d'argent. Ce travail introduit ensuite une approche à deux seuils, calibrée selon la procédure de Benjamini-Hochberg afin de garantir un taux de faux positifs contrôlé. Les expériences produites montrent que le transformeur est capable de produire des représentations générales qui mettent en évidence des schémas de blanchiment d'argent avec une supervision minimale des experts du domaine. En outre, elles démontrent comment exploiter ces représentations pour garantir un taux de faux positifs plus faible et contrôlé lors des investigations financières ultérieures.

Mots-clés. Réseaux de neurones, transformer, apprentissage contrastif, séries temporelles

Abstract. Financial institutions have an obligation to fight money laundering, but traditional rule-based approaches tend to be inefficient and expensive to operate due to a challenging compromise between detection rate and false positives. This research proposes a novel two-steps methodology for money laundering detection, leveraging on temporal data with a transformer. The first phase is focused on developing robust representations, independent of the available fraud labels, through contrastive learning. The second phase uses these representations to generate a money-laundering scoring. This work then introduces a two-threshold approach, calibrated following the Benjamini-Hochberg procedure to ensure a controlled false-positive rate. Experiments shows that the transformer is able to produce general representations that exhibit money-laundering patterns with minimal supervision from domain experts. Moreover, it shows how to leverage on these representations to ensure a lower and controlled false-positive rate for further financial investigations.

Keywords. Neural networks, transformer, contrastive learning, time series

1 Introduction

The United Nation (UNODC) defines money laundering as the processing of criminal proceeds to disguise their illegal origin [United Nation, 2025], making its detection particularly challenging. Indeed, perpetrators mimics legitimate financial behaviors, that can change quickly over time [Sinno, 2023]. This is a difficult challenge for detection, traditional approaches may become obsolete and financial patterns are more and more sophisticated.

Financial institutions operates under strict regulatory requirements for money laundering detection and suspicious activity reporting. Dedicated teams of financial analysts investigate potential cases. However, due to the high volume of customers and transactions in modern financial systems, these investigations cannot be entirely manual. Therefore, financial institutions have implemented so-called automated "rule-based systems" [Hossam, 2016 ; Laurenza, 2020], typically based on expert knowledge of past fraud patterns. Rule-based systems provide explicit audit and have some explanatory capabilities, which is an advantage in a regulatory context. However, it is difficult to maintain a high precision in such systems, some studies even report false positive rate from 95% to 98% [Lannoo, 2021]. There is opportunities of improvements.

Machine learning methods can address a lot of inefficiencies, and complement existing rule-based frameworks. One of the main contributions of this work on money laundering detection is to address it by training a transformer encoder through contrastive learning. The transformer allows to tackle time series data in an efficient way. Contrastive learning ensure a pre-training step that better handle low quality labels in AML, and that reduces the reliance on domain expertise. Moreover, the contrastive learning is performed with an innovative way of building positive and negative examples. Indeed, it reduces the amount of false negatives (negative example too similar to a reference observation in contrastive learning) by using an additional data set, to build a similarity criteria for the sampling. Finally, this work introduces of a two-thresholds methodology to better remove less suspicious observations while focusing on more critical observations, with a controlled false positive rate.

2 Representation Learning of financial time series

Real-life data set description Let us consider a data set $\mathcal{D} = \{x_0, x_1, x_2, \ldots, x_{N-1}\}$ of N observations ($x_i \in \mathbb{R}^{T \times d_{input}}, \forall i \in [0, N-1]$). In this work, the input observations x_is are time series of financial events with length T. Each time series $x_i = (x_{i,1}, \ldots, x_{i,T})$ is made of T vectors of dimension d_{input} describing T successive financial events. Each vector $x_{i,t}$ has both quantitative and qualitative features describing the event. These features can be the direction, the type, the date, and the amount of the financial transfer, for example.

In addition to these data $\mathcal{D}$, each individual also comes with complementary descriptors so that one can define a new set $\mathcal{D}' = \{x'_0, x'_1, x'_2, \ldots, x'_{N-1}\}$, where every $x'_i \in \mathbb{R}^{d_{comp}}$ is a vector of descriptors associated with the time series x_i. These features can be both categorical and numerical. The categorical ones include characteristics such as the legal form of a company, or the industry sector classification. The numerical features are mainly simple aggregated

metrics derived from the time series in $\mathcal{D}$.

Transformer trained with contrastive learning The transformer [Vaswani, 2017] is a powerful tool for processing time series [Wen, 2023]. Its architecture excels at capturing long-range dependencies, crucial in highlighting money laundering patterns. These patterns can be implemented over several months to be as discreet as possible. Moreover, the self-attention mechanism enables the encoder to focus on those patterns while building representations, filtering out less significant information in the time series, even in across long sequences.

A transformer encoder require a training objective in order to produce relevant representations of times series in $\mathcal{D}$. Contrastive learning [Le-Khac, 2020] is a flexible approach that allow the transformer encoder to train for this purpose. It is a self-supervised learning (SSL) method that exploits the intrinsic data structure to generate a training signal without manual labeling. This is done by creating a pseudo-supervision by leveraging relationships within the data to train the model to learn the semantic of unlabeled data. By doing this, contrastive learning can produce competitive results on a variety of tasks compared to supervised counterparts. Therefore, the transformer produces time series representations z, and minimizes the following InfoNCE loss:

$$\mathcal{L}_{InfoNCE}(\eta) = \mathbb{E}_{\mathcal{D},\mathcal{D}'}\left[-log\left(\frac{exp(\rho(z_+, z_{ref})/\mathcal{T})}{\sum_{z_j \in \mathcal{X}} exp(\rho(z_j, z_{ref})/\mathcal{T})}\right)\right] \tag{1}$$

$\mathcal{Z}_+ = \mathcal{Z}_+(z_{ref}) = \{z \mid z \text{ is a positive example with respect to } z_{ref}\}$
$z_+ = z_+(z_{ref}) \in \mathcal{Z}_+$
$\mathcal{Z}_- = \mathcal{Z}_-(z_{ref}) = \{z \mid z \text{ is a negative example with respect to } z_{ref}\}$
$z_- = z_-(z_{ref}) \in \mathcal{Z}_-$
$\mathcal{X} = \left\{z_+ \cup \left\{z_-^1, z_-^2, \cdots, z_-^{K-1}\right\} \text{ with } z_+ \in \mathcal{Z}_+, \left\{z_-^1, z_-^2, \cdots, z_-^{K-1}\right\} \in Z_-\right\}$
$\rho(\cdot)$ is a similarity measure, $\mathcal{T}$ is the temperature parameter

Positive and negative example selection For each (reference) observation, contrastive learning requires to sample one positive (i.e. similar) and K-1 negative (i.e. dissimilar) examples through a similarity criteria. While the data set $\mathcal{D}$ contains a lot of information, key in money laundering detection, the evaluation of financial time series similarity requires expert knowledge. On another hand, data set $\mathcal{D}'$ provides an insufficient amount of information to predict accurately money laundering. However, it contains descriptor vectors that can be used to build a similarity criteria, and allows to build positive and negative examples for any given reference observation. If two elements in $\mathcal{D}'$ are similar, their linked time series in $\mathcal{D}$ are expected to be similar. The proposition is therefore to sample the positive and negative examples using a similarity measure in $\mathcal{D}'$, implemented via a memory bank storing encoder's representation to enable efficient retrieval and a sampling diversity.

3 Money laundering detection

From representation learning to money-laundering detection Following representation learning, money laundering detection can be seen as a classification task requiring a second optimization. Let $\mathcal{Y} = \{0, 1\}$ be the set of possible labels (fraudsters vs non-fraudsters), with $\{y_0, y_1, y_2, \ldots, y_{N-1}\} \subset \mathcal{Y}$ the observed labels in the training set of length N. $\ell_c(\cdot)$ is a classification loss (e.g. binary cross-entropy), evaluating the detection performance. The first proposed approach fixes the transformer's weight θ, and consider the representations as static input features. Let $g_\mu(\cdot)$ be this new classifier of parameters μ, trained by minimizing the following loss:

$$\mathcal{L}_{classif-RP}(\mu) = \mathbb{E}_{\mathcal{D},\mathcal{Y}}\left[\ell_c\left(y, g_\mu\left(f_\theta(x)\right)\right)\right] \tag{2}$$

In this optimization, the optimal parameter values is noted $\hat{\mu}_{RP}$.

The other approach explored utilizes pre-training as a strategic initialization, by appending a classification head $g_\mu(\cdot)$ of parameters μ (e.g. a multi-layer perceptron). It directly connects the transformer encoder $f_\theta(\cdot)$ to class prediction. A joint optimization of both θ and μ is performed by minimizing the following loss:

$$\mathcal{L}_{classif-FT}(\theta, \mu) = \mathbb{E}_{\mathcal{D},\mathcal{Y}}\left[\ell_c(y, g_\mu(f_\theta(x)))\right] \tag{3}$$

In this second optimization, the optimal parameter values is noted $\left(\hat{\theta}_{FT}, \hat{\mu}_{FT}\right)$.

A two-thresholds approach The score produced by the resulting model can be interpreted as the probability of being a fraudster. The classical money laundering context comes with a huge imbalance between the two classes to be recovered: the dataset only contains a very small proportion of fraudsters. This strong imbalance suggests to use a two-thresholds strategy for tackling differently non-fraudsters and fraudsters. A first threshold T_l ("low threshold") is settled for removing the less suspicious observations from the list of observations triggered by alerts, and a second (larger) threshold T_h ("high threshold") is also designed for targeting the more suspicious observations for further investigations. As a consequence, analysts can save time by only focusing on observations that are most likely fraudsters. On the other side, a lot of observations can now be removed from the set of suspicious observations, avoiding any waste of time by further investigating on them.

Threshold calibration via Benjamini-Hochberg procedure For both problems a decision is taken for each account, while the proportion of wrong detections should be kept under control. Therefore the statistical performance is quantified by means of the False Discovery Rate (FDR) [Benjamini, 1995] especially designed to tackle such problems. In the present work, the calibration of the thresholds T_l and T_h is made by using a classical multiple testing procedure called the Benjamini–Hochberg (BH) procedure [Benjamini, 1995 ; Benjamini, 2021].

4 Experiments

The following figures isolate some experimental validations central to our theoretical methodology. Experiments are performed on real-life financial data to ensure a realistic evaluation, with anonymised data to guarantee privacy.

Representation learning evaluation through clustering Figure 1 demonstrates class homogeneity in the learned embedding space. The graph on the left shows that most of the fraudsters are concentrated on the right of the space. The graph on the right applies a K-Means clustering with two clusters, revealing one cluster aligning with the fraudster distribution observed on the left. This consistency validates the encoder's capacity to isolate illicit behavior without label supervision.

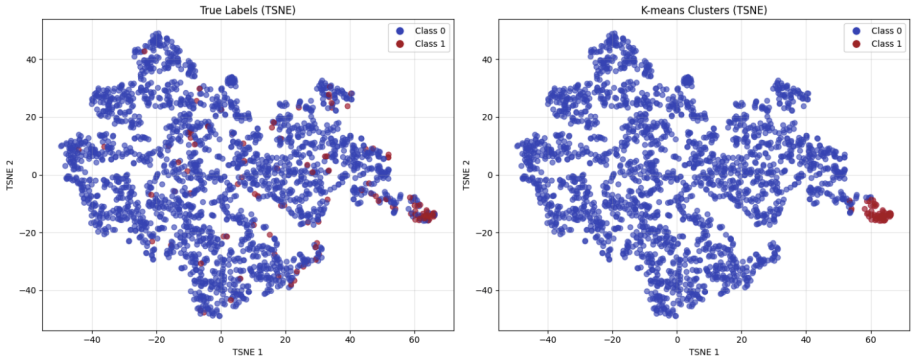


Figure 1: Embeddings of the transformer projected with t-SNE - Comparison between money-laundering labels (left panel) and learned clusters (right panel)

Distributions of fraudster and non-fraudster This work aims to differentiate fraudsters from non-fraudsters from the time series. After a representation has been learned in a latent space, a classifier is trained to produce a classification score, as described in the section 3. The figure 2 below shows the score learned by a logistic regression on top of the learned representations (that are fixed, cf equation 2). The two thresholds described in the section 3 can be learned from these results with a controled false-positive rate, thanks to the Benjamini-Hochberg procedure [Benjamini, 1995 ; Benjamini, 2021].

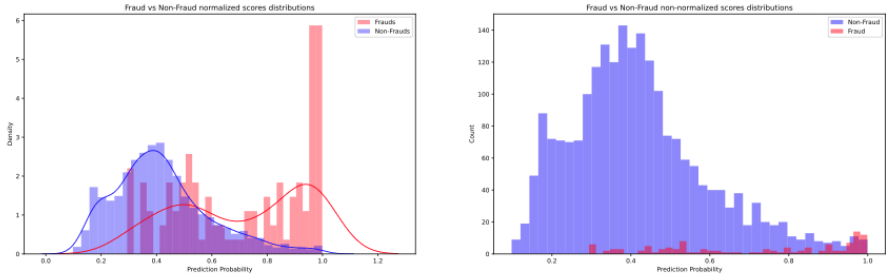


Figure 2: Normalized and non-normalized classification score distribution - Logistic Regression on contrastive learned representations

References

United Nation. (2025), Money laundering, proceeds of crime and the financing of terrorism.

Roy Majed Sinno, G. Yarrow Baldock, and Kimberly Gleason. (2023), The evolution of trade-based money laundering schemes: a regulatory dialectic perspective. *Journal of Finan- cial Crime*.

Tamer Hossam, Mohamed Zaki, Tarek Sobh, and Khaled Badran. (2016), Design of a monitor for detecting money laundering and terrorist financing. *Journal of Theoretical and Applied Information Technology, 85:425–436*.

Eleonora Laurenza, Luigi Bellomarini, and Emanuel Sallinger. (2020), Rule-based anti-money laundering in financial intelligence units: Experience and vision. 07 2020.

Karel Lannoo and Richard Parlour. (2021), Anti-money laundering in the eu: Time to get serious. *ceps task force report 28 jan 2021.*.

Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polo- sukhin Ashish Vaswani, Noam Shazeer. (2017), Attention is all you need. Advances in Neural Information Processing Systems 30 (NIPS 2017).

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. (2023), Transformers in time series: A survey. *In International Joint Conference on Artificial Intelligence(IJCAI)*.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. (2020), Contrastive representation learning: A framework and review. *IEEE Access, 8:193907–193934*.

Yoav Benjamini and Yosef Hochberg. (1995), Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1):289–300.

Yoav Benjamini and Daniel Yekutieli. (2001), The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics, 29(4):1165 – 1188.

Comparing and Proposing Methods for Estimating Bilateral Migration Flows

Thibault Laurent,[1] Gautier Debaud,[2] Paula Margaretic,[3] & Christine Thomas-Agnan[4]

[1] *Toulouse School of Economics, CNRS, France, Thibault.Laurent@tse-fr.eu*
[2] *GAEL, Université Grenoble Alpes, France, Gautier.Debaud@grenoble-inp.fr*
[3] *University of Chile, Chili, paumargaretic@gmail.com*
[4] *Toulouse School of Economics, France, Christine.Thomas@tse-fr.eu*

Résumé. Les données sur les flux migratoires entre paires de pays sont peu accessibles, malgré leur importance capitale pour les chercheurs — notamment en économie, sociologie et démographie — ainsi que pour l'élaboration de politiques publiques efficaces. Face à cette problématique, la seule solution disponible consiste à utiliser des estimations basées sur les données de stocks bilatéraux de migrants fournies par la Banque mondiale et les Nations Unies. Abel and Cohen (2019) ont recensé au moins cinq méthodes d'estimation possibles, mais aucune étude n'a encore effectué une comparaison exhaustive de ces méthodes. Dans cet article, nous évaluons et comparons ces approches à l'aide de simulations, tout en montrant que certains résultats peuvent être établis de manière théorique. Après avoir mis en évidence l'importance des ajustements démographiques pour chaque méthode, nous proposons une nouvelle approche d'estimation qui intègre le taux de retour dans la méthode de différence des stocks. Nos résultats montrent que cette méthode est comparable à la méthode Pseudo-Bayes, la plus couramment utilisée, et pourrait même offrir de meilleurs résultats si des données auxiliaires sur les pourcentages de retours migratoires entre paires de pays étaient disponibles.

Mots-clés. Données de migration, Méthodes d'estimation, Simulation, Statistique publique.

Abstract. Reliable data on international migration flows remain limited, despite their crucial importance for researchers—spanning fields such as economics, sociology, demography, and epidemiology—and for the formulation of effective public policies. Understanding the drivers of migration, predicting flow patterns, and assessing their consequences are therefore significant challenges. To tackle this issue, at least five synthetic estimators have been developed using bilateral migrant stock data from the World Bank and the United Nations (Abel and Cohen, 2019). However, comprehensive comparisons of these methods are scarce. In this paper, we evaluate and compare these approaches through simulations, while also demonstrating that some results can be theoretically established. After highlighting the importance of demographic adjustments across all methods and underscoring the need for a standardized framework, we introduce a new estimation method that combines stock differencing with migration rates, offering a balance of flexibility, simplicity, and accuracy. Our results suggest that this approach performs as well as the Pseudo-Bayes method and may even produce superior results if auxiliary data on return migration percentages between country pairs were available.

Keywords. Migration data, Estimation methods, Statistical simulation, Public statistics.

1 Introduction

Migration flows between pairs of countries present significant challenges due to various factors, including economic disparities, political instability, and social networks that influence migration patterns. Countries with large income gaps often experience high levels of outward migration, while those with better economic opportunities attract migrants. Political conflicts or environmental factors like climate change also play a crucial role in shaping migration trends. Climate shifts and the resulting disruptions are likely to have severe consequences for human populations, forcing some, particularly the most vulnerable, to migrate. Moreover, migration policies, border control mechanisms, and international agreements further complicate the dynamics between countries. Understanding these flows is essential for designing effective public policies that address the economic, social, and humanitarian needs of both sending and receiving nations.

Although some statistical offices and international organizations provide data on migration trends, these remain incomplete. For example, the UNHCR (United Nations High Commissioner for Refugees) publishes annual statistics on new asylum seekers and refugees. Meanwhile, the Organization for Economic Co-operation and Development (OECD) compiles annual data on labor migration, but only for its member countries, significantly limiting the scope of possible analyses.

Therefore, these sources cover only specific segments of global mobility, making it difficult to study migration trends as a whole, particularly in relation to climate change. Understanding how environmental factors influence migration decisions requires a more comprehensive dataset that captures global population movements. In the absence of direct data on migration flows, researchers must rely on estimates derived from bilateral migrant stock tables published by the World Bank and the United Nations.

Abel and Cohen (2019, 2022) introduced a series of methods for estimating bilateral migration flows that cover more than 200 countries and are detailed by sex. Although these estimators are widely used in migration studies, choosing the most appropriate method remains a challenge. The literature compares different estimation techniques, but there is no clear consensus on the best approach depending on the research question. Moreover, data selection can significantly impact study results, particularly when analyzing climate-induced migration. For example, Helbling et al. (2023) show that the use of one estimation method over another can lead to divergent conclusions regarding the relationship between climate change and migration flows.

In this study, we critically assess the methods proposed by Abel and Cohen (2019) and analyze their estimation errors using a systematic simulation approach. Our preliminary results show that errors vary depending on the type of migration, with some methods performing better in specific contexts. We also identify certain limitations in existing approaches, particularly the selective application of demographic adjustments, which are used in minimization and pseudo-Bayesian estimators but not incorporated into other methods. In response, we propose a new estimation approach combining stock differencing and migration rates, aiming to balance flexibility, simplicity, and accuracy.

2 Notations

We aim to estimate a flow matrix using stock data available through the United Nations. In this section, we first establish the framework of the study and the main notations used.

2.1 Migration flow matrices

Consider individuals who may choose to migrate between four countries A, B, C and D over a given period $[t_0, t_1]$. Since direct migration flow data are unavailable, our goal is to estimate the corresponding migration flow matrix F:

$$F = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} F_{AA} & F_{AB} & F_{AC} & F_{AD} \\ F_{BA} & F_{BB} & F_{BC} & F_{BD} \\ F_{CA} & F_{CB} & F_{CC} & F_{CD} \\ F_{DA} & F_{DB} & F_{DC} & F_{DD} \end{pmatrix} \end{array}, \tag{1}$$

where F_{ij} is the flow from the country of origin i, to the country of destination j, over the chosen period.

Every individual involved in these migration flows is necessarily born in one of the four countries A, B, C or D. Consequently, the total migration flow matrix F can be expressed as the sum of flow matrices by country of birth $F = F^A + F^B + F^C + F^D$, where F^k represents the migration flow matrix for individuals born in country F^k, $k \in \{A, B, C, D\}$. Moreover, their elements belong to one of the four categories shown in the diagram below.

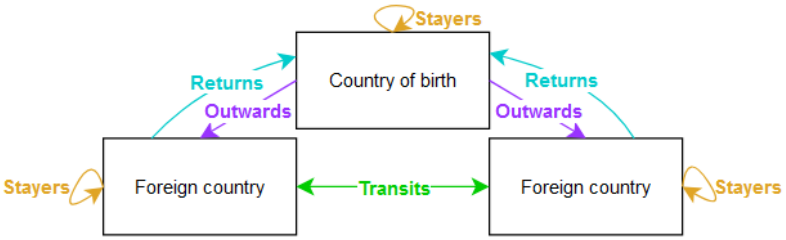


Figure 1: Decomposition of Flows by Country of Birth Into Four Categories

More precisely, for $(i, j, k) \in \{A, B, C, D\}^3$:

- The stayers, denoted Z^k_{ii}, represent individuals born in k who remained in country i throughout the period considered.
- The outwards migrants, denoted O^k_{kj}, represent individuals born in k who migrated to another country j during the period considered.
- The returnees, denoted R^k_{ik}, represent individuals born in k who previously lived in country i and returned to k over the period considered.

- The transit migrants, denoted T_{ij}^{k}, represent individuals born in k who moved from country $i \neq k$ to country $j \neq k$ during the period under consideration.

Using this decomposition, we can express the general form of the flow matrices by country of birth as follows:

$$F^{A} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} Z_{AA}^{A} & O_{AB}^{A} & O_{AC}^{A} & O_{AD}^{A} \\ R_{BA}^{A} & Z_{BB}^{A} & T_{BC}^{A} & T_{BD}^{A} \\ R_{CA}^{A} & T_{CB}^{A} & Z_{CC}^{A} & T_{CD}^{A} \\ R_{DA}^{A} & T_{DB}^{A} & T_{DC}^{A} & Z_{DD}^{A} \end{pmatrix} \end{array} \quad F^{B} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} Z_{AA}^{B} & R_{AB}^{B} & T_{AC}^{B} & T_{AD}^{B} \\ O_{BA}^{B} & Z_{BB}^{B} & O_{BC}^{B} & O_{BD}^{B} \\ T_{CA}^{B} & R_{CB}^{B} & Z_{CC}^{B} & T_{CD}^{B} \\ T_{DA}^{B} & R_{DB}^{B} & T_{DC}^{B} & Z_{DD}^{B} \end{pmatrix} \end{array}$$

$$F^{C} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} Z_{AA}^{C} & T_{AB}^{C} & R_{AC}^{C} & T_{AD}^{C} \\ T_{BA}^{C} & Z_{BB}^{C} & R_{BC}^{C} & T_{BD}^{C} \\ O_{CA}^{C} & O_{CB}^{C} & Z_{CC}^{C} & O_{CD}^{C} \\ T_{DA}^{C} & T_{DB}^{C} & R_{DC}^{C} & Z_{DD}^{C} \end{pmatrix} \end{array} \quad F^{D} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} Z_{AA}^{D} & T_{AB}^{D} & T_{AC}^{D} & R_{AD}^{D} \\ T_{BA}^{D} & Z_{BB}^{D} & T_{BC}^{D} & R_{BD}^{D} \\ T_{CA}^{D} & T_{CB}^{D} & Z_{CC}^{D} & R_{CD}^{D} \\ O_{DA}^{D} & O_{DB}^{D} & O_{DC}^{D} & Z_{DD}^{D} \end{pmatrix} \end{array}$$

As with the global flow matrix, terms in the same row have the same country of origin, those in the same column the same country of destination. The country of birth is added as a superscript. Naturally, departures (O), returns (R) and transits (T) will not be placed in the same way depending on the country of birth of the individuals observed.

All these flows can be linked with the following system of equations:

$$F_{ij} = \begin{cases} F_{ij}^{i} + F_{ij}^{j} + \sum_{k \neq \{i,j\}} F_{ij}^{k} = O_{ij}^{i} + R_{ij}^{j} + \sum_{k \neq \{i,j\}} T_{ij}^{k} & \text{if } i \neq j \\ \sum_{k} F_{ii}^{k} = \sum_{k} Z_{ii}^{k} & \text{if } i = j \end{cases}$$

2.2 Matrices of migrant stocks

To estimate flows, all methods rely on stock matrices. For four countries, these matrices take the following form:

$$S(t) = \begin{array}{c} \\ A \\ B \\ C \\ D \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} S_{A}^{A}(t) & S_{B}^{A}(t) & S_{C}^{A}(t) & S_{D}^{A}(t) \\ S_{A}^{B}(t) & S_{B}^{B}(t) & S_{C}^{B}(t) & S_{D}^{B}(t) \\ S_{A}^{C}(t) & S_{B}^{C}(t) & S_{C}^{C}(t) & S_{D}^{C}(t) \\ S_{A}^{D}(t) & S_{B}^{D}(t) & S_{C}^{D}(t) & S_{D}^{D}(t) \\ TR_{A}(t) & TR_{B}(t) & TR_{C}(t) & TR_{D}(t) \end{pmatrix} \end{array} \begin{array}{c} \\ TB^{A}(t) \\ TB^{B}(t) \\ TB^{C}(t) \\ TB^{D}(t) \\ N(t) \end{array} \tag{2}$$

where $S_{j}^{i}(t)$ is the number of people born in i and living in j at time t, $TB^{i}(t)$ is the number of people born in i who are still alive at time t, $TR_{j}(t)$ is the number of people living in j at time t, and $N(t)$ is the total number of people in the system at time t.

It is possible to express stocks using stayers, outwards migrants, returnees, and transit migrants. We apply this approach at two points in time, t_0 and t_1. The stocks of individuals

born in i and living in $j \neq i$ at time t_0 can be obtained by summing the individuals who: remain in j at t_1, return to i at t_1, or transit to another country $k \notin i, j$ at t_1. The stocks of individuals born in i and living in i at time t_0 are obtained by summing the individuals who: remain in i at t_1 or leave for another country $k \neq i$ at t_1. Thus, we can write :

$$S_j^i(t_0) = \begin{cases} Z_{jj}^i + R_{ji}^i + \sum_{k \neq \{i,j\}} T_{jk}^i & \text{if } i \neq j \\ Z_{ii}^i + \sum_{k \neq i} O_{ik}^i & \text{if } i = j \end{cases}.$$

The stocks of people born in i and living in $j \neq i$ at time t_1 are obtained by summing the people who: are in j at t_0 and remain there, are in i at t_0 and leave for j, or are in another country $k \notin i, j$ at t_0 and transit to j. The stocks of people born in i and living in i at time t_1 are obtained by summing the people who: are in i at t_0 and remain there, or are in another country $k \neq i$ at t_0 and return to i. Thus, we can write:

$$S_j^i(t_1) = \begin{cases} Z_{jj}^i + O_{ij}^i + \sum_{k \neq \{i,j\}} T_{kj}^i & \text{if } i \neq j \\ Z_{ii}^i + \sum_{k \neq i} R_{ki}^i & \text{if } i = j \end{cases}.$$

Using the previous decompositions of flows and stocks, we can show that, on the one hand, the elements of the k^{th} row of $S(t_0)$ are the row margins of the matrix F^k, and on the other hand, the elements of the k^{th} column of $S(t_1)$ are the column margins of the matrix F^k. For example, if we add the margins of F^A:

$$F^A = \begin{array}{c} \\ A \\ B \\ C \\ D \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} Z_{AA}^A & O_{AB}^A & O_{AC}^A & O_{AD}^A \\ R_{BA}^A & Z_{BB}^A & T_{BC}^A & T_{BD}^A \\ R_{CA}^A & T_{CB}^A & Z_{CC}^A & T_{CD}^A \\ R_{DA}^A & T_{DB}^A & T_{DC}^A & Z_{DD}^A \end{pmatrix} \\ \begin{array}{cccc} S_A^A(t_1) & S_B^A(t_1) & S_C^A(t_1) & S_D^A(t_1) \end{array} \end{array} \begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{l} \\ \\ \\ \\ S_A^A(t_0) \\ S_B^A(t_0) \\ S_C^A(t_0) \\ S_D^A(t_0) \end{array}$$

3 Estimation method

Building on the relationships between flows and stocks outlined above, several methods can be employed to estimate migration flows. To the best of our knowledge, five primary estimation methods are commonly used in migration studies, categorized into three groups: Stock differencing methods, migration rate-based methods and Model-based methods. We implement a new method that combines stock differencing with migration rates. Due to space constraints, we illustrate only the method based on the difference of stock matrices. It relies on the difference $(\Delta S)_j^i = S_j^i(t_1) - S_j^i(t_0)$ in the size of migrant stocks at the start and end of a given period to estimate the corresponding migration flow. This difference can also be viewed as the net result of inflows and outflows in j for individuals born in i.

$$(\Delta S)^i_j = \begin{cases} (O^i_{ij} + \sum_{k\neq\{i,j\}} T^i_{kj}) - (R^i_{ji} + \sum_{k\neq\{i,j\}} T^i_{jk}) = \text{Inflows}^i_j - \text{Outflows}^i_j, & \text{if } i \neq j \\ \sum_{k\neq i}(R^i_{ki} - O^i_{ik}), & \text{if } i = j \end{cases} \tag{3}$$

where Inflows^i_j refers to all individuals born in i who move to j, whether from i itself or from other countries, and Outflows^i_j refers to all individuals born in i who leave j, whether to return to i or to transit through other locations. In the *drop negative* method, the flow from i to j ($i \neq j$) is estimated by calculating the difference in stocks over the period of interest. Any negative differences are treated as zero flows and are set to zero:

$$\hat{F}^{\text{drop}}_{ij} = (\Delta S)^i_j \times \mathbb{1}_{(\Delta S)^i_j > 0}. \tag{4}$$

The residuals $F_{ij} - \hat{F}_{ij}$ can be expressed in terms of stayers, outflows, returns, and transits. After simplification, the following cases emerge:

$$\epsilon^{\text{drop}}_{ij} = \begin{cases} R^j_{ij} + R^i_{ji} + \sum_{k\neq\{i,j\}} T^k_{ij} - \sum_{k\neq\{i,j\}}(T^i_{kj} - T^i_{jk}) & \text{pour } (\Delta S)^i_j > 0 \\ O^i_{ij} + R^j_{ij} + \sum_{k\neq\{i,j\}} T^k_{ij} & \text{pour } (\Delta S)^i_j \leq 0 \end{cases}$$

4 Simulation results

We run 100 simulations using parallel programming and graphically compare results:

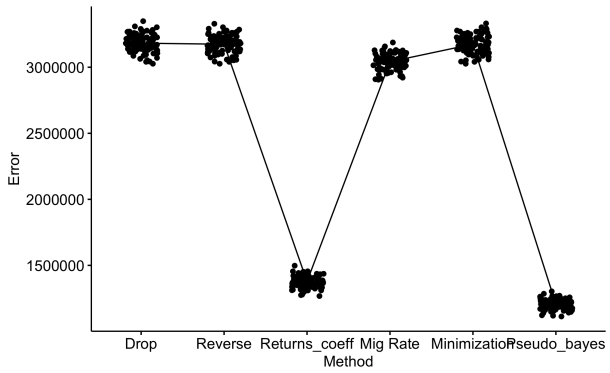


Figure 2: Boxplot MSE and Overestimation/Underestimation

Looking at the MSE boxplot, it is clear that the *pseudo-Bayes* method performs the best, with the *return coefficients* method trailing closely behind. The other methods appear fairly similar to one another, but with much larger errors compared to these top two. Methods like *Drop*, *reverse*, *migration rates*, and *minimization* tend to underestimate almost all flows, while *return coefficients* and *pseudo-Bayes* may either overestimate or underestimate them. Interestingly, the flows where *return coefficients* and *pseudo-Bayes* perform less well tend to be those they overestimate—this is not particularly surprising, given that these flows are relatively small in this initial example.

References

Abel, G. J. and Cohen, J. E. (2019). Bilateral international migration flow estimates for 200 countries. *Scientific data*, 6(1):82.

Abel, G. J. and Cohen, J. E. (2022). Bilateral international migration flow estimates updated and refined by sex. *Scientific Data*, 9(1):1–11.

Helbling, M., Rybski, D., Schewe, J., Siedentop, S., Glockmann, M., Heider, B., Jones, B., Meierrieks, D., Rikani, A., and Stroms, P. (2023). Measuring the effect of climate change on migration flows: Limitations of existing data and analytical frameworks. *PLOS Climate*, 2(1):e0000078.

Statistique spatiale

Lead Prediction Error Analysis in 2D-RCAR Random Fields

Ouerdia Arezki[1], Sylvie Viguier[1], Sébastien Pinel[2]

[1] *LAMPS, IUT de Carcassonne, Université de Perpignan Via Domitia*
Email : ouerdia.arezki@univ-perp.fr, viguier@univ-perp.fr
[2] *CEFREM, IUT de Carcassonne, Université de Perpignan Via Domitia*
Email : sebastien.pinel@univ-perp.fr

Abstract. In this paper, we deal with the prediction of the autoregressive model with indexed random coefficients in two dimensions of order $(p_1, p_2) \in \mathbb{N}^2$, $2D - RCAR(p_1, p_2)$ in short. We have derived an expression for the asymptotic mean squared error an (h_1, h_2) step ahead forecast of predicted value from general nonlinear random field models with random coefficients more precisely, for general two-dimensionally indexed random coefficients autoregressive of order (p_1, p_2) when the parameters are known. The predictive performance is analyzed using MATLAB simulations. The results demonstrate that our predictor provides robust multi-step forecasts, with accuracy influenced by lead time and model parameters.

Keywords. Spatial Statistics, 2D-RCAR Model, Nonlinear Prediction, lead Prediction Error

1 Introduction and Problem Formulation

Spatial prediction methods have seen significant advancements due to high demand in fields like agriculture, climatology, ecology, econometrics, geology, medical studies, oil prospecting, and water pollution analysis, where data is typically available at specific locations [1, 2, 3, 5, 9]. The increasing complexity of spatial data and the growing need for accurate predictive tools have driven researchers to develop advanced models that go beyond traditional linear frameworks. Linear models, while straightforward and computationally efficient, often fail to capture the intricate dependencies and nonlinear relationships present in spatial datasets. This limitation has underscored the necessity for more sophisticated approaches that can accommodate these complexities.

Two-dimensional indexed random coefficient autoregressive models (2D-RCAR) are a generalization of 2D-AR models. The introduction of 2D-RCAR models is motivated by the fact that the fixed coefficients hypothesis for the 2D-AR Model is not very realistic. Indeed, the autoregressive models with coefficients assumed to be not constant, but subject to random perturbations allow accurate handling of the non-homogeneity of a process and ensure a faithful interpretation of real-valued spatial phenomena [2, 4, 6, 7, 8]. Bibi. et al [4] realized a well-structured framework for analyzing the $2D - RCAR(p_1, p_2)$, $(p_1, p_2) \in \mathbb{N}^2$ models. These models allow us to exploit all the analytical tools and formulate new results. We borrow the following definition, which gives the main analytical features of the class of $2D - RCAR(p_1, p_2)$ models.

Definition 1.1 ([4]). *Autoregressive valued random field* $\{X_{s,t}, (s,t) \in \mathbb{Z}^2\}$ *defined in the probability space* (Ω, F, P) *is called two-dimensional indexed random coefficient autoregressive model with order* $(p_1, p_2) \in \mathbb{N}^2$ *denoted by* $2D - RCAR(p_1, p_2)$ *if it is a solution of the following stochastic equation*

$$X_{s,t} = \sum_{\substack{i=0 \\ (i,j)\neq(0,0)}}^{p_1} \sum_{j=0}^{p_2} (\xi_{i,j} + \xi_{s,t}^{(i,j)}) X_{s-i,t-j} + \varepsilon_{s,t}, \quad (s,t) \in \mathbb{Z}^2. \tag{1}$$

where

(i) *The innovation process* $\{\varepsilon_{s,t}, (s,t) \in \mathbb{Z}^2\}$ *is an independent and identically distributed (i.i.d) sequence of random fields with mean zero and variance* σ^2.

(ii) $\underline{\xi} = (\xi_{i,j}, (i,j) \in \Im)^\top$ *is the vector of fixed coefficients, representing the global dependence structure. It is a real-valued vector of constants, where for any positive integers* p_1 *and* p_2*, the set of indices is given by:* $\Im = \{(i,j) \mid 0 \leq i \leq p_1,\ 0 \leq j \leq p_2,\ (i,j) \neq (0,0)\}$.

These coefficients define the spatial interactions between observations, ensuring a structured dependence across the field.

(iii) $\underline{\xi}_{s,t} = (\xi_{s,t}^{(i,j)}; (i,j) \in \Im, (s,t) \in \mathbb{Z}^2)$ *is taken with respect to the lexicographic order. Then, the perturbation process* $(\underline{\xi}_{s,t}, (s,t) \in \mathbb{Z}^2)$ *is a strictly stationary process of independent vectors of random fields with zero mean and variance:* $var(\underline{\xi}_{s,t}) = E(\underline{\xi}_{s,t}\underline{\xi}^\top{}_{s,t}) = \Lambda$*, which is independent of:* $\{\varepsilon_{s,t}, (s,t) \in \mathbb{Z}^2\}$.

The perturbations $\xi_{s,t}^{(i,j)}$ *introduce stochastic components to model local variations and heterogeneities. They allow the model to capture local variations, making it more suitable for heterogeneous spatial processes. Observations vary based on local random coefficients, enabling flexibility in spatial modeling.*

The matrix $\Lambda = Var(\xi_{s,t})$ *captures the dependencies between the different components of the perturbations* $\xi_{s,t}$*. If* $\xi_{s,t} = (\xi_{s,t}^{(i,j)})$*, then* Λ *is a* $p_1 \cdot p_2 \times p_1 \cdot p_2$ *matrix, where each element* $\Lambda_{(i,j),(k,l)}$ *represents the covariance between* $\xi_{s,t}^{(i,j)}$ *and* $\xi_{s,t}^{(k,l)}$*. As a covariance matrix, it is symmetric:* $\Lambda_{(i,j),(k,l)} = \Lambda_{(k,l),(i,j)}$.

The matrix Λ *describes the variance and covariance structure of the perturbations, ensuring the model can represent complex relationships between random coefficients. A typical matrix* Λ *for a 2D-RCAR model of order* (p_1, p_2) *has the following structure:*

$$\Lambda = \begin{pmatrix} Var(\xi_{s,t}^{(1,1)}) & Cov(\xi_{s,t}^{(1,1)}, \xi_{s,t}^{(1,2)}) & \cdots & Cov(\xi_{s,t}^{(1,1)}, \xi_{s,t}^{(p_1,p_2)}) \\ Cov(\xi_{s,t}^{(1,2)}, \xi_{s,t}^{(1,1)}) & Var(\xi_{s,t}^{(1,2)}) & \cdots & Cov(\xi_{s,t}^{(1,2)}, \xi_{s,t}^{(p_1,p_2)}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\xi_{s,t}^{(p_1,p_2)}, \xi_{s,t}^{(1,1)}) & Cov(\xi_{s,t}^{(p_1,p_2)}, \xi_{s,t}^{(1,2)}) & \cdots & Var(\xi_{s,t}^{(p_1,p_2)}) \end{pmatrix}.$$

(iv) *The variance of* $\varepsilon_{s,t}$*,* $\sigma^2 \geq c_1 > 0$ *, while the smallest eigenvalue of* Λ *is bounded below by* c_2*, where* c_1 *and* $c_2 > 0$ *are both arbitrarily small*

Let $W_{st} = K_{m^2-1} vec(Y_{s,t-1} Y_{s,t-1}^\top)$*, where* $Y_{s,t-1} = (X_{k,l}; (1,1) \leq (k,l)\ < (m,m))^\top$.

To address the challenges of forecasting spatio-temporal data, the predictor for future values $X(n+h_1, m+h_2)$ leverages the inherent structure of the 2D-RCAR model. This predictor accounts for the relationships encoded in past observations and integrates randomness in the coefficients, offering a nuanced approach to modeling complex dependencies. By recursively utilizing past observations, the proposed methodology captures the dynamic nature of spatio-temporal processes, making it robust for diverse applications. The recursive nature of the prediction framework is particularly advantageous in handling multi-step-ahead forecasting. Each step incorporates the evolving

relationships between spatial locations and temporal intervals, ensuring that the predictor remains adaptable to both local and global patterns in the data. This adaptability is critical in scenarios such as environmental monitoring, where spatial and temporal changes often exhibit nonlinearity and randomness. Now, we define the vector form of the model (1) by

$$X_{s,t} = (\Psi + \Xi_{s,t})X_{s,t-1} + \mathcal{E}_{s,t}, \tag{2}$$

where

$$\mathbf{X}_{s,t} = \begin{pmatrix} X_{s,t} \\ X_{s-1,t} \\ \vdots \\ X_{s-p_1,t} \\ X_{s,t-1} \\ \vdots \\ X_{s,t-p_2} \end{pmatrix}_{p_1 \cdot p_2 \times 1}, \mathcal{E}_{s,t} = \begin{pmatrix} \varepsilon_{s,t} \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{p_1 \cdot p_2 \times 1}, \Psi = \begin{pmatrix} \psi_{1,1} & \cdots & \psi_{1,p_2} \\ \vdots & \ddots & \vdots \\ \psi_{p_1,1} & \cdots & \psi_{p_1,p_2} \end{pmatrix}_{p_1 \times p_2}, \tag{3}$$

$$\Xi_{s,t} = \begin{pmatrix} \xi_{s,t}^{(1,1)} & \cdots & \xi_{s,t}^{(1,p_2)} \\ \vdots & \ddots & \vdots \\ \xi_{s,t}^{(p_1,1)} & \cdots & \xi_{s,t}^{(p_1,p_2)} \end{pmatrix}_{p_1 \times p_2}.$$

This compact expression will be used in the next section for deriving asymptotic mean squared prediction error at lead period (h_1, h_2).

Under the assumption of stationarity on the random field [12] developed a framework for computing the predictors of random fields by extending some works of stationary time series cited in their bibliography, such as Pourahmadi (1992), Cheng and Pourahmadi (1997). Inspired by [12] we consider the problem of finding the autoregressive representation for all (h_1, h_2)-step ahead linear predictors, $(h_1, h_2) \geq (0, 0)$ of $X(s + h_1, t + h_2)$ based on the strict past of X at the point (s, t). Unlike the 1-D case, there is no natural order definition in the 2-D domain. However, the totally ordered NSHP support is a favorable type of support in the sense that it yields a natural extension to the 1-D results. This fact has been shown by [10, 11].

Definition 1.2 ([10]). *We call a nonsymmetrical half-plan past (NSHP) any subset S of $\mathbb{Z}^2$ satisfying*

1. *S stable under addition*
2. *$S \cup -S = \mathbb{Z}^2$*
3. *$S \cap -S = \{(0, 0)\}$.*

The NSHP support S considered in this paper results from the total-order definition (4a). The total order is defined as a 'raster-scan': top-to-bottom column-after-column from left to right. In mathematical terms: The order $<$ is defined by

$$(s_1, t_1) < (s_2, t_2) \text{ iff } (s_1, t_1) \in \{(k, l)|k = s_2, l < t_2\} \cup \{(k, l)|k < s_2, -\infty < l < \infty\}. \tag{4a}$$

and the order $\leq$ is naturally defined by

$$(s_1, t_1) \leq (s_2, t_2) \text{ iff } (s_1, t_1) < (s_2, t_2) \text{ or } (s_1, t_1) = (s_2, t_2). \tag{4b}$$

which coincides with the lexicographic order.

2 Main Result: Asymptotic Mean Squared PredictionError at Lead Period (h_1, h_2)

We consider the problem of predicting future values of the process $X_{s,t}$, defined on a two-dimensional lattice. The forecast targets locations $(s + h_1, t + h_2)$ lying beyond the observed region S, which are essential for capturing the spatiotemporal dynamics of the model. In this setting, we assume that the parameter vector $\underline{\xi} = (\xi_{i,j},\, (i,j) \in \Im)^\top$ is known. The uncertainty in prediction is thus solely due to the random innovations and their propagation through the autoregressive structure. This analysis provides an explicit expression for the forecast error, which reflects both the effect of innovation propagation and the variability induced by the random coefficients. It also defines a theoretical lower bound on forecast accuracy, serving as a benchmark for later comparisons when the parameters are estimated. The following result formalizes this setting.

Theorem 2.1. *Let $X_{s,t}$ be a strictly stationary and $\mathcal{F}_{s,t}$-measurable process satisfying equation* (1)*, under assumptions* (**i**) *and* (**iv**)*.*

Then, for any $(s + h_1, t + h_2) \in S^c$, where S^c denotes the complement of the index set S, the (h_1, h_2)-step ahead predictor of $X_{s+h_1,t+h_2}$, with $(h_1, h_2) \geq (0, 0)$ and $(h_1, h_2) \neq (0, 0)$, based on the available past values in S, is given by

$$\widehat{\mathbf{X}}_{s+h_1,t+h_2} = \Psi^{h_1 \times h_2 - 1} \mathbf{X}_{s,t}.$$

The corresponding forecast error satisfies

$$\widehat{\mathbf{X}}_{s+h_1,t+h_2} - \mathbf{X}_{s+h_1,t+h_2} = -\left(\mathcal{M}_{h_1,h_2} - \Psi^{h_1 \times h_2 - 1}\right) \mathbf{X}_{s,t} - \sum_{\substack{k=0 \\ (k,l)\neq(h_1,h_2)}}^{h_1} \sum_{l=0}^{h_2} \mathcal{M}_{k,l} \mathcal{E}_{s+h_1-k,t+h_2-l}.$$

As a result, the asymptotic mean squared prediction error (AMSE) at horizon (h_1, h_2) is given by

$$\mathbb{E}\left[\left(\widehat{\mathbf{X}}_{s+h_1,t+h_2} - \mathbf{X}_{s+h_1,t+h_2}\right)^2\right] = \operatorname{tr}\left\{\mathbf{V} \cdot \mathbb{E}\left[\left(\mathcal{M}_{h_1,h_2} - \Psi^{h_1 \times h_2 - 1}\right) G \left(\mathcal{M}_{h_1,h_2} - \Psi^{h_1 \times h_2 - 1}\right)^\top\right]\right\} + \sigma_a^2 \sum_{\substack{k=0 \\ (k,l)\neq(h_1,h_2)}}^{h_1} \sum_{l=0}^{h_2} \operatorname{tr}\left\{\mathbb{E}\left(\mathcal{M}_{k,l} G \mathcal{M}_{k,l}^\top G\right)\right\}.$$

The precise definitions of the matrices $\Psi^{h_1 \times h_2 - 1}$, $\mathcal{M}_{k,l}$, $\mathbf{V}$, and G will be provided in the subsequent sections.

Let suppose that $\underline{\xi} = (\xi_{i,j}, (i,j) \in \Im)^\top$ *is known.* Then, from (2) we can write

$$\begin{aligned} X(s + h_1, t + h_2) &= \mathcal{E}_{s+h_1,t+h_2} + (\Psi + \Xi_{s+h_1,t+h_2})\mathcal{E}_{s+h_1,t+h_2-1} \\ &\quad + (\Psi + \Xi_{s+h_1,t+h_2})(\Psi + \Xi_{s+h_1,t+h_2-1})\mathcal{E}_{s+h_1,t+h_2-2} + \dots \\ &\quad + (\Psi + \Xi_{s+h_1,t+h_2}) \dots (\Psi + \Xi_{s+h_1,t+2})\mathcal{E}_{s+h_1,t+1} \\ &\quad + (\Psi + \Xi_{s+h_1,t+h_2}) \dots (\Psi + \Xi_{s+h_1,t+1}) X_{s,t} \qquad (5) \\ &= \sum_{\substack{k=0 \\ (k,l)\neq(h_1,h_2)}}^{h_1} \sum_{l=0}^{h_2} \mathcal{M}_{k,l} \mathcal{E}_{s+h_1-k,t+h_2-l} + \mathcal{M}_{h_1,h_2} \mathbf{X}_{s,t}. \end{aligned}$$

where

$$\mathcal{M}_{0,0} = 1$$
$$\mathcal{M}_{k,l} = (\Psi + \Xi_{s+h_1,t+h_2})\,(\Psi + \Xi_{s+h_1,t+h_2-1,m+h_2-1}) \ldots (\underline{\Psi} + \Xi_{s+h_1,t+h_2-k,m+h_2-l})$$

Now, substituting the future values of $\varepsilon_{s+h_1,t+h_2}$ and $\left(\xi_{s,t}^{(i,j)}\right)$; $(i,j) \in \Im, (s,t) \in \mathbb{Z}^2$ of (5) by their expectations, which are zero, we write the forecast value of $X_{s+h_1,t+h_2}$ at lead period (h_1, h_2) as

$$\hat{\mathbf{X}}_{s+h_1,t+h_2} = \Psi^{h_1 \times h_2 - 1} X_{s,t}. \tag{6}$$

Hence the forecast error is

$$\hat{\mathbf{X}}_{s+h_1,t+h_2} - \mathbf{X}_{s+h_1,t+h_2} = -(\mathcal{M}_{h_1,h_2} - \Psi^{h_1 \times h_2 - 1})\mathbf{X}_{s,t} - \sum_{\substack{k=0 \\ (k,l)\neq(h_1,h_2)}}^{h_1} \sum_{l=0}^{h_2} \mathcal{M}_{k,l}\mathcal{E}_{s+h_1-k,t+h_2-l}. \tag{7}$$

Defining $\mathbf{e}^t = \left(1, 0, \ldots, 0\right)^t$, we may write $\mathbf{X}_{s+h_1,t+h_2}$ as

$$\mathbf{X}_{s+h_1,t+h_2} = \mathbf{e}^t \mathbf{X}_{s+h_1,t+h_2}. \tag{8}$$

Hence, the mean squared forecast error at lead period $(h-1, h-2)$ is

$$\begin{aligned} E\left(\hat{\mathbf{X}}_{s+h_1,t+h_2} - \mathbf{X}_{s+h_1,t+h_2}\right)^2 &= E[\mathbf{e}^\top \left(\hat{\mathbf{X}}_{s+h_1,t+h_2} - \mathbf{X}_{s+h_1,t+h_2}\right)\left(\hat{\mathbf{X}}_{s+h_1,t+h_2} - \mathbf{X}_{s+h_1,t+h_2}\right)^\top \mathbf{e}] \\ &= E\left[\mathbf{e}^\top (\mathcal{M}_{h_1,h_2} - \Psi^{h_1 \times h_2 - 1})\mathbf{X}_{s,t}\mathbf{X}_{s,t}^\top (\mathcal{M}_{h_1,h_2} - \Psi^{h_1 \times h_2 - 1})\mathbf{e}\right] \\ &\quad + \sum_{\substack{k=0 \\ (k,l)\neq(h_1,h_2)}}^{h_1} \sum_{l=0}^{h_2} E\left(\mathbf{e}^\top \mathcal{M}_{k,l}\mathcal{E}_{s+h_1-k,t+h_2-l}\mathcal{M}_{k,l}^\top \mathcal{E}_{s+h_1-k,t+h_2-l}^\top \mathbf{e}\right) \\ &= \mathrm{tr}\left[\mathbf{V} E\{(\mathcal{M}_{h_1,h_2} - \Psi^{h_1 \times h_2 - 1})G(\mathcal{M}_{h_1,h_2} - \Psi^{h_1 \times h_2 - 1})^\top\}\right] \\ &\quad + \sigma_a^2 \sum_{\substack{k=0 \\ (k,l)\neq(h_1,h_2)}}^{h_1} \sum_{l=0}^{h_2} \mathrm{tr} E\left(\mathcal{M}_{k,l} G \mathcal{M}_{k,l}^\top G\right). \end{aligned} \tag{9}$$

where

$$G = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix}^{p_1^2 \times p_2^2}$$

3 Numerical Example

To illustrate the practical application of our proposed linear predictor, we consider a numerical example where we simulate and predict values for a 2D-RCAR process given by:

$$X_{s,t} = \sum_{i=0}^{2} \sum_{j=0}^{2} \left(\xi_{i,j} + \xi_{s,t}(i,j)\right) X_{s-i,t-j} + \varepsilon_{s,t}, \quad (s,t) \in \mathbb{Z}^2,$$

where the fixed coefficients $\xi_{i,j}$ were defined as: $\xi = \begin{bmatrix} 0.5 & 0.3 \\ 0.2 & 0.1 \end{bmatrix}$. The implementation is performed using MATLAB, we generated on a 50×50 spatial grid. Perturbations $\xi_{s,t}(i,j)$ were independent and identically distributed (i.i.d.) random variables with zero mean and variance $\sigma_\xi^2 = 0.05$. The noise process $\varepsilon_{s,t}$ was i.i.d. with zero mean and variance $\sigma_\varepsilon^2 = 0.1$. Observations were generated recursively using the model equations, starting with initial values set to zero. The top-left subplot in Figure 1 represents the actual state, while the other subplots illustrate the predicted states for leads $(1,1)$, $(2,2)$, and $(10,10)$. The color scale indicates the magnitude of the states, highlighting the accuracy of predictions across different lead periods. Figure 2 shows the errors in state estimation

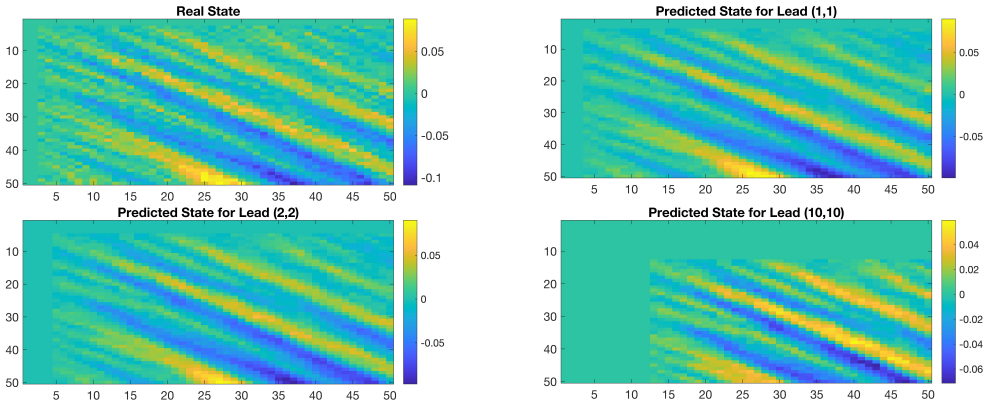


Figure 1: Actual and Predicted States for Different Leads

over time. The errors remain within acceptable bounds for most time steps, indicating the robust performance of the estimation model. Figure 3 presents a boxplot of the empirical AMSE for

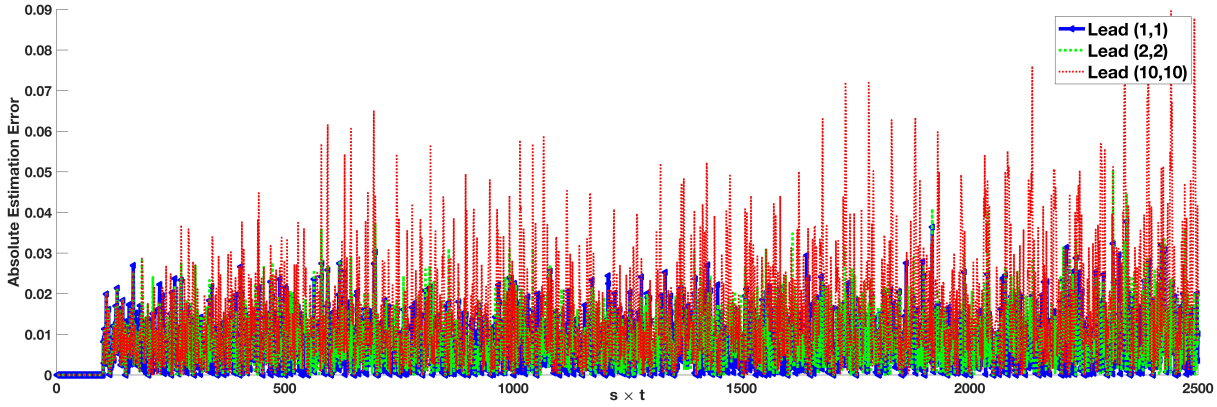


Figure 2: Estimation Errors Between Actual and Predicted States

different lead lengths $(1,1)$, $(2,2)$, and $(10,10)$. The results show that AMSE increases with lead length, which is expected due to the accumulation of prediction uncertainty over longer horizons. Notably, the variability in AMSE across simulations is minimal for each lead, indicating the model's consistency and robustness. This trade-off between lead length and prediction accuracy underscores the importance of selecting an appropriate lead based on the desired application.

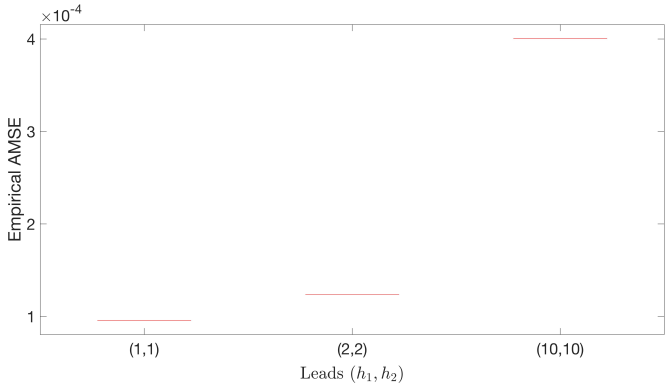


Figure 3: Boxplot of Empirical AMSE Across Different Leads

4 Conclusion & Prospects

In this paper, we have derived an expression for the asymptotic mean squared error of predicted value for more than one step ahead from general nonlinear random field models with random coefficients more precisely, for 2D-RCAR (p_1, p_2) models. We have considered the case when the parameters are known, as future work, we aim to treat the problem when the parameters $\underline{\xi} = (\xi(i,j), (i,j) \in \Im)^T$ are estimated.

References

[1] Cressie, N. (1993). *Statistics for Spatial Data*. Wiley.

[2] Wikle, C. K. (1999). Dimension Reduction in Spatial Dynamic Models. *Journal of the American Statistical Association*.

[3] Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society, Series B*.

[4] Bibi, H., et al. (2015). A Framework for Analyzing 2D-RCAR Models.

[5] Ripley, B. D. (1981). *Spatial Statistics*. Wiley.

[6] Guyon, X. (1995). *Random Fields on a Network: Modeling, Statistics, and Applications*. Springer.

[7] Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC.

[8] Saidi, A., Hamaz, A., & Arezki, O. (2022). Estimation in nonlinear random fields models of autoregressive type with random parameters. Communications in Statistics - Theory and Methods, 53(1), 294–309.

[9] Arezki, O., & Hamaz, A. (2020). On linear prediction for stationary random fields with non-symmetrical half-plane past. Communications in Statistics - Theory and Methods, 51(15), 5298–5309. https://doi.org/10.1080/03610926.2020.1837880

[10] Helson, H., & Lowdenslager, D, Prediction theory and Fourier series in several variables, Acta Mathematica, 99(1), 165-202 (1958).

[11] Helson, H., & Lowdenslager, D, Prediction theory and Fourier series in several variables. II, Acta Mathematica, 106(3-4), 175-213 (1961).

[12] Kohli, P, & Pourahmadi, M, Some prediction problems for stationary random fields with quarter-plane past. Journal of Multivariate Analysis, 127, 112-125 (2014).

Statistique des points critiques d'un champ gaussien

Julien Chevallier [1] & Jean François Coeurjolly [2]

[1] *Université Grenoble Alpes, France, julien.chevallier1@univ-grenoble-alpes.fr*
[2] *Université Grenoble Alpes, France, jean-francois.coeurjolly@univ-grenoble-alpes.fr*

Résumé. Nous proposons de faire une revue de la littérature (probabiliste) récente consacrée aux propriétés des points critiques d'un champ gaussien vus comme la réalisation d'un processus ponctuel. En particulier, l'article de Azaïs et Delmas (2022) propose l'étude la plus complète (à notre connaissance) à propos de l'attraction/repulsion à faible portée, et l'article de Azaïs et al. (2024) propose un Théorème de la limite centrale pour le nombre de points critiques, multivarié en l'indice des-dits points critiques.

Nous agrémentons cette revue d'une étude numérique non triviale du comportement à faible portée (pour illustrer les résultats théoriques), mais aussi à plus grande portée. De plus, nous étendons le Théorème de la limite centrale au cas des fonctionnelles linéaires et bilinéaires du processus ponctuel. Enfin, nous montrons un résultat de consistence de l'approche par simulation.

Mots-clés. Champ gaussien, processus ponctuel

Abstract. A review of the recent (probabilistic) litterature dealing with critical points of gaussian random fields (seen as a realization of a point process) is proposed. In particular, the article of Azaïs et Delmas (2022) gives the most complete study (up to our knowledge) of their low range attractive/repulsive behavior. Moreover, the article of Azaïs et al. (2024) gives a multivariate Central Limit Theorem for these critical points (*multivariate* means that the indices of the critical points are taken into account).

This review is complemented with a non trivial numerical study of the low range behavior (illustrating theorical results) but also at a higher range. Moreover, the CLT is extended to linear and bilinear functionals of the point process. Finally, the convergence of the simulated critical points to the correct distribution is provided.

Keywords. Gaussian field, point process

1 Gaussian random fields

Let $\mathbf{X} = \{X(s), s \in \mathbb{R}^d\}$ be a real-valued Gaussian random field on $\mathbb{R}^d$ ($d \geq 1$). We assume in particular that $\mathbf{X}$ is centered, stationary and isotropic and we denote by $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ its covariance function, thus invariant by translation and rotation. We abuse notation by still denoting c the stationary covariance, and propose to denote by c_1 and c_2 the following functions, defined for any $s, t \in \mathbb{R}^d$

$$c(s,t) = \mathbb{E}\{X(s)X(t)\} = c(s-t) = c_1(\|s-t\|) = c_2(\|s-t\|^2) \tag{1}$$

where $\|\cdot\|$ stands for the Euclidean norm. For instance, the Gaussian covariance given by $c(s,t) = \exp\{-(s-t)^2/\varphi^2\}$ (for some parameter $\varphi > 0$) corresponds to $c_1(r) = \exp(-r^2/\varphi^2)$ and $c_2(r) = \exp(-r/\varphi^2)$. The notation c_2 may appear useless or unusual but some of the results, in particular from Azaïs et Delmas (2022) are better expressed using c_2 instead of c_1.

Two models are thoroughly discussed:

1. The **Matérn correlation** function parameterized by a scale parameter $\varphi > 0$, a regularity parameter $\nu > 0$ and given for any $r > 0$ by

$$c_1(r) = c_2(r^2) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{r\sqrt{2\nu}}{\varphi}\right)^{\nu} K_\nu \left(\frac{r\sqrt{2\nu}}{\varphi}\right) \tag{2}$$

   where K_ν is the modified Bessel function of the second kind. The parameter ν is related to the regularity of the field: it can be shown that it is at least $\lceil \nu \rceil - 1$ times continuously differentiable in the mean square and almost sure sense.

2. The **Gaussian Random Wave Model** (RWM). Its covariance function writes, for $d \geq 1$

$$c_1(r) = c_2(r^2) = \Gamma\left(\frac{d}{2}\right) \left(\frac{r\sqrt{d}}{2\varphi}\right)^{-(d/2-1)} J_{d/2-1}\left(\frac{r\sqrt{d}}{\varphi}\right), \tag{3}$$

   where J_ν stands for the Bessel function with order ν. The RWM is infinitely continuously differentiable in the mean square and almost sure sense.

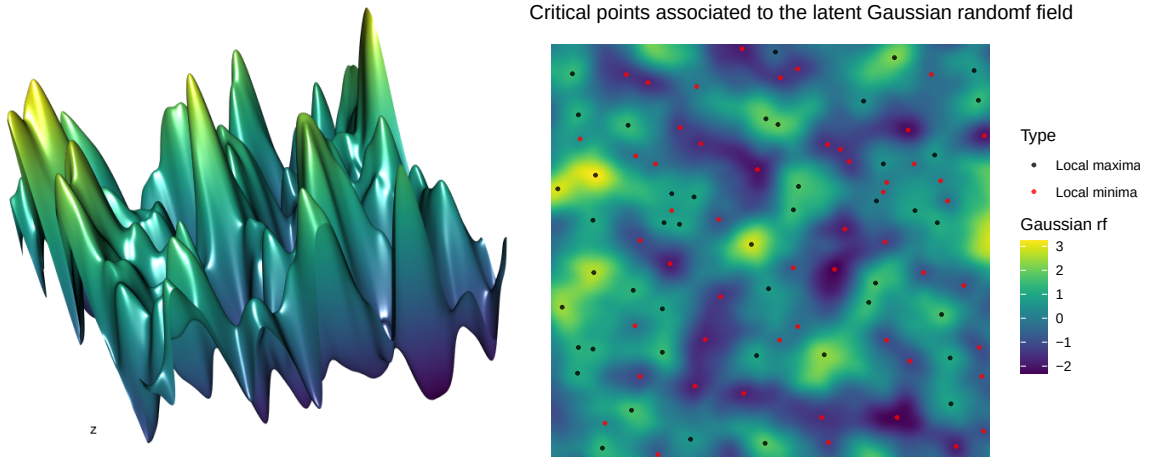


Figure 1: (Left) Simulation of a Gaussian random field $\mathbf{X}$ in $[0,1]^2$ with Gaussian correlation function; (Right) The set of black (resp. red points) is a realization of $\mathbf{Y}_{\mathcal{L}}$ with $\mathcal{L} = d$ (resp. with $\mathcal{L} = 0$) that is the set of local maxima (resp. local minima).

2 Point processes

We consider simple spatial point processes in $\mathbb{R}^d$, that is a random locally finite subset $\mathbf{Y}$ of a Borel set $S \subseteq \mathbb{R}^d$, $d \geq 1$. Furthermore, the point processes of interest here are also stationary and isotropic. The notation $\mathbf{Y} \cap B$ stands for the restriction of $\mathbf{Y}$ to a set $B \subseteq \mathbb{R}^d$ and we let $|B|$ denote the volume of any bounded $B \subset \mathbb{R}^d$. Local finiteness of $\mathbf{Y}$ means that $\mathbf{Y} \cap B$ is finite almost surely (a.s.), that is the number of points $N(B) := \#(\mathbf{Y} \cap B)$ of $\mathbf{Y} \cap B$ is finite a.s., whenever B is bounded.

By stationarity, the intensity ρ is such that $\mathbb{E}[N(B)] = \rho|B|$ for all $B \subseteq \mathbb{R}^d$ and the pair correlation function g is given for any pairwise distinct $s, t \in \mathbb{R}^d$ by

$$g(s,t) = \lim_{\varepsilon \to 0} \frac{\mathbb{E}[N(B(s,\varepsilon))N(B(t,\varepsilon))]}{\mathbb{E}[N(B(s,\varepsilon))]\,\mathbb{E}[N(B(t,\varepsilon))]}.$$

By stationarity and isotropy, g is a function of $\|s-t\|$ only and we make the abuse of notation, $g(s,t) = g(\|s-t\|)$. The pair correlation is a measure of departure of the model to the Poisson point process. Poisson point processes are the reference models which generate points without any interaction. In particular, $g \equiv 1$ for a Poisson process. Hence, when $g(r) < 1$ (resp. $g(r) > 1$) two points at distance r are less likely (resp. more likely) to appear than under a Poisson point process. A point process is said to be repulsive or clustered (resp. purely repulsive or purely clustered) if $g(0) < 1$ or $g(0) > 1$ (resp. $g(r) < 1$ or $g(r) > 1$ for any r).

3 Critical points of a gaussian random field

3.1 Intensity and pair correlation function

Here is the definition relating the two previous sections.

Definition 1 (Critical point process)**.** *Let $\mathbf{X}$ be a centered, stationary and isotropic Gaussian random field on $\mathbb{R}^d$ ($d \geq 1$), with unit variance and at least twice continuously differentiable. For any $\ell = 0, \ldots, d$, we define by $\mathbf{Y}_\ell$ the point process obtained from the critical points with index ℓ of $\mathbf{X}$ and more generally the point processes obtained from the critical points with indices $\mathcal{L} \subseteq \{0, \ldots, d\}$ as follows*

$$\mathbf{Y}_\ell = \left\{ t \in \mathbb{R}^d : X'(t) = 0, \iota\{X''(t)\} = \ell \right\} \quad \text{and} \quad \mathbf{Y}_{\mathcal{L}} = \cup_{\ell \in \mathcal{L}} \mathbf{Y}_\ell. \tag{4}$$

By definition, the point processes $\mathbf{Y}_0$ and $\mathbf{Y}_d$ (for which $\mathcal{L} = \{0\}$ or $\{d\}$) correspond to the random sets of local minima and local maxima of $\mathbf{X}$ while the set of all critical points will be denoted $\mathbf{Y}_{0:d}$ which corresponds to $\mathcal{L} = \{0, 1, \ldots, d\}$.

The notation $N_{\mathcal{L}}(B)$ stands for the number of critical points with indices $\mathcal{L}$ which fall into a bounded domain $B \subset \mathbb{R}^d$. The intensity of $\mathbf{Y}_{\mathcal{L}}$ is denoted $\rho_{\mathcal{L}}$ and the cross pair correlation function is defined, for any $\mathcal{L}, \mathcal{L}' \subseteq \mathbb{R}^d$, as

$$g_{\mathcal{L},\mathcal{L}'}(s,t) = g_{\mathcal{L},\mathcal{L}'}(\|t-s\|) = \lim_{\varepsilon \to 0} \frac{\mathbb{E}[N_{\mathcal{L}}(B(s,\varepsilon))N_{\mathcal{L}'}(B(t,\varepsilon))]}{\mathbb{E}[N_{\mathcal{L}}(B(s,\varepsilon))]\,\mathbb{E}[N_{\mathcal{L}'}(B(t,\varepsilon))]}. \tag{5}$$

The paper Azaïs and Delmas (2022) contains the following results as soon as $\mathbf{X}$ is almost surely C^2:

- expression of $\rho_{\mathcal{L}}$ (explicit for $d \leq 4$) in terms of the derivatives $c_2'(0)$ and $c_2^{(2)}(0)$,
- asymptotic behavior of $g_{\mathcal{L},\mathcal{L}'}(r)$ as $r \to 0$.

The ingredients in their proof even yields an expression of $g_{\mathcal{L},\mathcal{L}'}(r)$ for all $r > 0$ by means of Kac-Rice formula. It is not explicit but can be estimated via Monte-Carlo procedure (see Figure 2).

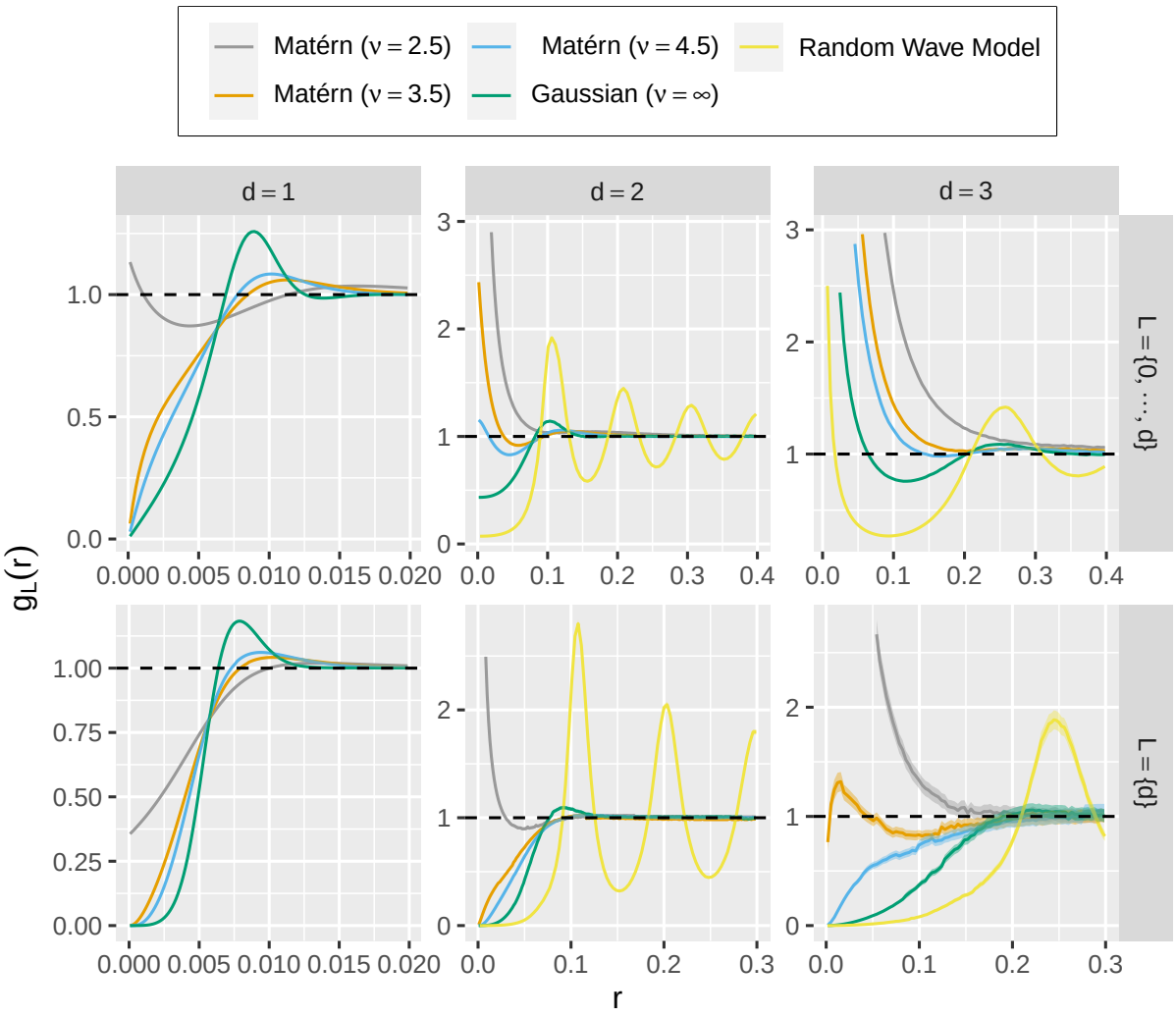


Figure 2: Monte-Carlo estimation (with $B = 10^7$ simulations) of pair correlation function for critical points with indices $\mathcal{L} = \{0, \ldots, d\}$ and $\mathcal{L} = \{d\}$ (for $d = 1, 2, 3$) based on Matérn random fields with regularity parameter $\nu = 2.5, 3.5, 4.5, \infty$ and on the Random Wave Model (for $d > 1$). The scale parameter φ is set such that $\rho_{\mathcal{L}} = 100$.

3.2 Statistics of (bi)linear functionals

Considering $\mathbf{Y}_{\mathcal{L}}$ as a stochastic model, it is of great interest to have asymptotic results for functionals of point processes when the amount of information increases. For that we consider the increasing domain asymptotic and assume, to ease the presentation, that the critical point process is observed in the sequence of bounded domains $W_n = [-n/2, n/2]^d$ (with volume n^d). Let $\mathcal{L}, \mathcal{L}' \subseteq \{0, \dots, d\}$. Let also $\phi_{1,n} : \mathbb{R}^d \to \mathbb{R}$ and $\phi_{2,n} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be two sequences of measurable (test) functions. We consider the functionals $\Phi_{1,n}$ and $\Phi_{2,n}$ defined by

$$\Phi_{1,n} = \Phi_{1,n}(\phi_{1,n}, \mathbf{Y}_{\mathcal{L}}, W_n) = \sum_{t \in \mathbf{Y}_{\mathcal{L}} \cap W_n} \phi_{1,n}(t) \tag{6}$$

$$\Phi_{2,n} = \Phi_{2,n}(\phi_{2,n}, \mathbf{Y}_{\mathcal{L}}, \mathbf{Y}_{\mathcal{L}'}, W_n) = \overset{\neq}{\sum_{s \in \mathbf{Y}_{\mathcal{L}} \cap W_n, t \in \mathbf{Y}_{\mathcal{L}'} \cap W_n}} \phi_{2,n}(s, t), \tag{7}$$

where the sign $\neq$ means that the sum is defined for any pairwise distinct points s, t. Statistics of the form (6) (resp. (7)) are called linear statistics (resp. bilinear statistics). On the one side, linear ones capture the natural unbiased estimator of the intensity parameter $\rho_{\mathcal{L}}$ or the estimation of a d-dimensional integral using quadrature points obtained as realizations of a point process. On the other side, bilinear ones capture the standard estimator of Ripley's K function.

We are now in position to state our main result.

Theorem 1. *Let $\mathcal{L}, \mathcal{L}' \subseteq \{1, \dots, d\}$. Under technical assumptions on $\mathbf{X}$ and normalizing assumptions on $\phi_{1,n}$ and $\phi_{2,n}$, we have*

$$n^{d/2} \left(\Phi_{1,n} - \mathbb{E}\left[\Phi_{1,n}\right]\right) \xrightarrow{\mathcal{D}} N(0, \sigma_1^2) \tag{8}$$

and

$$n^{d/2} \left(\Phi_{2,n} - \mathbb{E}\left[\Phi_{2,n}\right]\right) \xrightarrow{\mathcal{D}} N(0, \sigma_2^2), \tag{9}$$

where σ_1^2 and σ_2^2 are covariances invovling the chaos expansion of $\mathbf{X}$.

Bibliographie

J.-M. Azaïs and C. Delmas (2022), Mean number and correlation function of critical points of isotropic Gaussian fields and some results on GOE random matrices. *Stochastic Processes and their Applications*, 150:411–445.

J.-M. Azaïs, F. Dalmao, and C. Delmas (2024), Multivariate CLT for critical points, arXiv:2401.09117.

Estimation de la régularité d'un champ gaussien

Alexandre Makhlouf [1]

[1] *Avignon Université, LMA, UPR2151, F-84000 Avignon, France,*
alexandre.makhlouf@univ-avignon.fr

Résumé. On présente un estimateur $\widehat{r}_n$ de la régularité pour un champ gaussien. Cet estimateur est construit à partir d'une unique réalisation du champ observé sur une grille uniforme. Il généralise l'estimateur construit par Blanke et Vial (2014) pour des processus gaussiens. La construction de l'estimateur repose sur l'étude de la vitesse de convergence des variations quadratiques généralisées. Sous certaines hypothèses de régularité, on obtient un estimateur fortement consistent. Parmi les familles de champs aléatoires qui vérifient ces hypothèses, on retrouve notamment les champs browniens fractionnaires et les champs de Matérn. Estimer cette régularité permet d'estimer celle de la fonction de covariance et ainsi, par exemple, d'améliorer le calibrage d'un modèle pour l'interpolation. Le travail présenté s'inscrit dans un objectif plus grand : construire un interpolateur polynômial par morceaux, dont le degré local est donné par la régularité du champ. Cela permet entre autres de s'affranchir d'une contrainte technique liée aux méthodes de krigeage : faire un choix de modèle de covariance.

Mots-clés. Estimateur, Régularité, Champ aléatoire

Abstract. We present an estimator $\widehat{r}_n$ of the regularity for a Gaussian random field. This estimator is constructed from a single realisation of the observed field on a uniform grid. It generalises the estimator constructed by Blanke and Vial (2014) for Gaussian random processes. Its construction is based on the study of the rate of convergence of the generalised quadratic variations. under certain regularity assumptions, we obtain a strongly consistent estimator. The families of random fields that verify these assumptions include fractional Brownian fields and Matérn fields. Estimating this regularity also enables us to estimate the regularity of the covariance function and therefore, for example, improve the fitting of a model for interpolation. The work presented here is part of a larger project : building a piecewise polynomial interpolator, with a local degree given by the regularity of the field. Among other things, this makes it possible to overcome a technical constraint associated with kriging methods : choosing a covariance model.

Keywords. Estimator, Mean-square regularity, Random field

1 Introduction

1.1 État de l'art

L'estimation de la régularité, et plus particulièrement de la régularité des champs lisses, est un sujet peu étudié dans la littérature. Blanke et Vial (2014) ont construit un estimateur de cette régularité pour des processus gaussiens. Le processus doit vérifier des conditions de régularité, notamment une condition de stationnarité locale (voir Berman (1974)). Loh (2015) a construit un estimateur pour des champs gaussiens, stationnaires et isotropes, dont la régularité est régie par un paramètre $\nu \in (0, 2)$. Sa construction repose sur des variations quadratiques du second ordre. Loh, Sun et Wen (2021) ont construit un estimateur pour des champs gaussiens stationnaires dont la fonction de covariance est donnée par une fonction de Matérn isotrope de paramètre de régularité ν. À notre connaissance, ce sont les seuls travaux qui traitent de l'estimation de la régularité pour des processus et des champs aléatoires réguliers.

On présente ici un nouvel estimateur de la régularité pour des champs gaussiens qui généralise celui construit par Blanke et Vial (2014) dans le cadre des processus gaussiens. On ne prescrit pas à la covariance du champ aléatoire une forme explicite et on n'impose pas au champ aléatoire d'être stationnaire.

1.2 Cadre d'étude

Dans ce résumé, on utilise les conventions d'écriture multi-indicielles. Les éléments en gras représentent des vecteurs.

Soient $\mathcal{D} = [0, 1]^2$, $\mathfrak{D}$ un voisinage de $\mathcal{D}$ et n un entier strictement positif. Considérons $X = \{X(\mathbf{s})\}_{\mathbf{s}\in\mathfrak{D}}$ un champ gaussien à valeurs réelles. On suppose connaitre une réalisation du champ X sur la grille $\left\{\frac{\mathbf{i}}{n},\, \mathbf{i} \in [\![\mathbf{0}, \mathbf{n}]\!]\right\}$, où $[\![\mathbf{0}, \mathbf{n}]\!] = [\![0, n]\!]^2$ et $[\![0, n]\!]$ est l'ensemble des entiers compris entre 0 et n.

Soit r_0 un entier naturel. On suppose en plus que le champ aléatoire X est r_0-fois continument différentiable en moyenne quadratique sur $\mathfrak{D}$, *i.e.* que le champ X admet des dérivées partielles en moyenne quadratique d'ordre $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$, avec $\alpha_1 + \alpha_2 \leq r_0$, sur tout $\mathfrak{D}$ et continues en moyenne quadratique sur $\mathfrak{D}$.

2 Estimation de la régularité

L'objectif est de construire un estimateur $\widehat{r}_n$ de l'ordre r_0 de différentiabilité du champ aléatoire X, à l'aide de ses observations. Pour ce faire on va utiliser les variations quadratiques généralisées. Les variations quadratiques d'un processus ou d'un champ aléatoire non régulier sont couramment utilisées pour estimer le paramètre de régularité höldérienne, comme par

exemple dans Istas et Lang (1997) pour des processus ou dans Benassi, Cohen et Istas (2002) pour des champs.

2.1 Variation quadratique généralisée

Soit $r \leq n$ un entier strictement positif. Considérons c_r une famille de $(r+1)^2$ réels vérifiant

$$\begin{cases} \sum_{\mathbf{i=0}}^{\mathbf{r}} c_r(\mathbf{i})\mathbf{i}^{\boldsymbol{\alpha}} = 0, & \text{si } |\boldsymbol{\alpha}| < r, \\ \sum_{\mathbf{i=0}}^{\mathbf{r}} c_r(\mathbf{i})\mathbf{i}^{\boldsymbol{\alpha}} \neq 0, & \text{si } |\boldsymbol{\alpha}| = r, \end{cases}$$

où $\mathbf{i}^{\boldsymbol{\alpha}} = i_1^{\alpha_1} \times i_2^{\alpha_2}$ et $|\boldsymbol{\alpha}| = \alpha_1 + \alpha_2$. La famille de réels c_r a des propriétés similaires à ce que l'on appelle un filtre dans la littérature, comme par exemple dans Coeurjolly (2001) ou dans Biermé, Bonami et Leon (2011).

Pour tout $\mathbf{k} \in [\![\mathbf{0}, \mathbf{n} - \mathbf{r}]\!]$, on définit l'accroissement généralisé du champ X au point $\frac{\mathbf{k}}{n}$ par

$$\Delta_n^r X(\mathbf{k}) = \sum_{\mathbf{i=0}}^{\mathbf{r}} c_r(\mathbf{i}) X(\tfrac{\mathbf{k}+\mathbf{i}}{n})$$

et on considère la variation quadratique généralisée de X suivante

$$V_n^r(X) = \frac{1}{n^2} \sum_{\mathbf{k=0}}^{\mathbf{n-r}} \left(\Delta_n^r X(\mathbf{k})\right)^2.$$

Théorème 1
Soit r_0 un entier positif. Sous certaines hypothèses de régularité sur le champ X, il existe une constante $H_0 \in (0,1)$ tel que :

1. *pour tout $r_0 \geq 1$ et pour tout $r \in [\![1, r_0]\!]$, il existe une constante $D(r) > 0$ telle que*

$$n^{2r}\mathbb{E}\left[V_n^r(X)\right] \xrightarrow[n\to+\infty]{} D(r),$$

2. *pour tout entier $r > r_0$, il existe une constante $C(r, r_0) \geq 0$ telle que*

$$n^{2(r_0+H_0)}\mathbb{E}\left[V_n^r(X)\right] \xrightarrow[n\to+\infty]{} C(r, r_0).$$

Le point remarquable dans ce résultat est le changement de la vitesse de convergence dès lors que r dépasse r_0 et que $C(r, r_0) \neq 0$. Le réel H_0 est lié à la régularité höldérienne des dernières dérivées partielles du champ aléatoire X.

2.2 Estimateur

La construction de l'estimateur est basé sur la remarque précédente. On va comparer la variation quadratique généralisée $V_n^r(X)$ à une suite $(\ell_n)_{n\in\mathbb{N}}$ qui a le comportement suivant. Considérons $(\ell_n)_{n\in\mathbb{N}}$ une suite vérifiant

$$\begin{cases} n^{2H_0-2}\ell_n \xrightarrow[n\to+\infty]{} 0, \\ n^{2H_0}\ell_n \xrightarrow[n\to+\infty]{} +\infty. \end{cases}$$

Considérons aussi une suite d'entiers $(b_n)_{n\in\mathbb{N}}$ telle que $2 \leq b_n \leq n$ pour tout n et $b_n \to +\infty$ quand n tend vers $+\infty$. On définit l'estimateur $\widehat{r}_n$ de la façon suivante :

$$\widehat{r}_n = \min\left\{r \in [\![2, b_n]\!] \,:\, n^{2(r-1)}V_n^r(X) \geq \ell_n\right\} - 2, \tag{1}$$

et où l'on impose à $\widehat{r}_n$ une valeur non-entière arbitraire si l'ensemble introduit en (1) est vide.

Théorème 2
Soit r_0 un entier positif. Sous certaines hypothèses de régularité sur le champ gaussien X et en supposant que $C(r, r_0) \neq 0$, il existe une constante $c > 0$ telle que

$$\mathbb{P}\left(\{\widehat{r}_n \neq r_0\}\right) = \mathcal{O}\left(\exp\left(-c\left(n\mathbb{1}_{(0,\frac{1}{2})}(H_0) + \tfrac{n}{\ln(n)}\mathbb{1}_{\{\frac{1}{2}\}}(H_0) + n^{2(1-H_0)}\mathbb{1}_{(\frac{1}{2},1)}(H_0)\right)\right)\right).$$

En particulier, presque-sûrement, pour n assez grand, on a $\widehat{r}_n = r_0$.

Bibliographie

Berman, S. M. (1974), Sojourns and Extremes of Gaussian Processes, *The Annals of Probability*, 2(6), pp. 999-1026

Benassi, A., Cohen, S. et Istas, J. (2002), Identification and properties of real harmonizable fractional Lévy motions, *Bernoulli*, 8(1), pp. 97–115.

Biermé, H., Bonami, A. et Leon, J-R. (2011), Central limit theorems and quadratic variations in terms of spectral density, *Electronic Journal of Probability*, 16, pp. 362-395.

Blanke, D. et Vial, C. (2014), Global smoothness estimation of a Gaussian process from general sequence designs, *Electronic Journal of Statistics*, 8, pp. 1152–1187.

Coeurjolly, J-F. (2001), Estimating the parameters of a fractional Brownian motion by discrete variations of its sample paths, *Statistical Inference for Stochastic Processes*, 4(2), pp. 199-227.

Istas, J. et Lang, G. (1997), Quadratic variations and estimation of the local Hölder index of a gaussian process, *Annales de l'I.H.P. Probabilités et statistiques*, 33(4), pp. 407-436.

Loh, W-L. (2015), Estimating the smoothness of a Gaussian random field from irregularly spaced data via higher-order quadratic variations, *The Annals of Statistics*, 43(6), pp. 2766-2794.

Loh, W-L., Sun, S. et Wen, J. (2021), On fixed-domain asymptotics, parameter estimatioin and isotropic gaussian randoim fields with Matérn covariance functions, *The Annals of Statistics*, 49(6), pp. 3127-3152.

Spatio-temporal random fields on meshed surfaces defined from advection-diffusion SPDEs

Mike Pereira [1] & Lucia Clarotto [2]

[1] *Centre for geosciences and geoengineering, Mines Paris, PSL University, France, mike.pereira@minesparis.psl.eu*
[2] *UMR MIA Paris-Saclay, AgroParisTech, INRAE, Université Paris-Saclay, France, lucia.clarotto@agroparistech.fr*

Résumé. Dans ce travail, nous introduisons une classe de processus gaussiens pour modéliser des données spatio-temporelles sur des surfaces (maillées) en présence d'advection et de diffusion dans les données. Suivant l'approche SPDE largement utilisée en statistiques spatio-temporelles, nous considérons une classe d'équations quz dérivées partielles d'advection-diffusion stochatiques définis sur des manifolds riemanniens fermés, compacts et orientables de dimension 2, et leur discrétisation à l'aide d'une approche de Galerkin. Nous montrons comment cette approche permet de proposer des algorithmes scalables pour la simulation et la prédiction de champs aléatoires gaussiens solutions de l'EDPS discrétisée.

Mots-clés. Processus gaussien, Statistiques spatio-temporelles, Variété riemannienne, EDPS.

Abstract. In this work, we introduce class of Gaussian processes to model spatio-temporal data on (meshed) surfaces in the presence of advection and diffusion in the data. Following the so-called SPDE modeling approach widely used in spatio-temporal statistics, we consider a class of advection-diffusion stochastic partial differential equations defined on smooth compact Riemannian manifolds of dimension 2, and their discretization using a Galerkin approach. We show how this approach allows to easily propose scalable algorithms for the simulation and prediction of Gaussian random fields that are solutions of the discretized SPDE.

Keywords. Gaussian process, Spatio-temporal statistics, Riemannian manifold, SPDE

1 Introduction

Geostatistics is a branch of statistics that focuses on modeling data distributed in space and time. Within the geostatistical paradigm, the observed variable is modeled as (a realization of) a space-time Gaussian random field (GRF), so that the mere characterization of its mean and covariance functions suffices to fully describe its statistical properties. In particular, these two functions are chosen to mimic the spatio-temporal variability and structure observed in the data Wackernagel (2003). This probabilistic framework has many advantages. On the one hand, it allows to perform simulations of random fields with the same

spatio-temporal structure as the one observed in the data, and predictions of at unobserved locations. On the other hand, uncertainties can be quantified both on the variable behavior at unobserved locations (using so-called conditional simulations) or on the model parameters (through Bayesian approaches) (Chiles and Delfiner; 2009).

Of particular interest in this project is the setting where the spatial domain on which the data lie is not Euclidean, but rather represents a surface. For instance, this is the case when dealing fMRI data in neuroimaging applications, in which case the data lie on the cortical surface (ie. the surface of the brain): see eg. Mejia et al. (2020). This is also the case when considering global data in environmental applications, for which the data lie on a sphere representing our planet (see eg. Rayner et al. (2020)), and on this surface transport phenomena (due to winds and currents for instance) can affect the structure of the data.. Hence the main motivations of this work: proposing models for spatio-temporal GRFs flexible enough to represent complex patterns of correlations in data lying on arbitrary surfaces, and but simple enough so that numerically efficient algorithms for their inference, simulation and prediction can be derived.

Since we consider GRFs, tasks such as sampling from (un)-conditional distributions, predictions (through conditional expectations),and likelihood-based inference can all be performed by solving linear systems or adequately factorizing covariance matrices of the field Wackernagel (2003). Hence, the most straightforward (and classical) approach to spatio-temporal geostatistical modeling consists in fitting valid space-time covariance functions on the data, so that these covariance matrices may be built. Nonetheless, the covariance-based approach has two main drawbacks. First, the matrix factorizations required in sampling, prediction and inference tasks have a complexity that scales as the cube of the number of observations and/or target points, thus making them unfeasible when this last number is large. To circumvent this, simplifying assumptions on the covariance model must be made, such as the separability of space and time dependencies or stationarity. This in turn may result in a lack of realism of the model. Secondly, since they rely on Euclidean or arc-length distances, most of the covariance models available in the literature are restricted to the setting where the spatial domain is either Euclidean or the sphere . Hence, they hardly generalize to other surfaces.

In this work, we consider GRFs based on the definition and discretization of an advection-diffusion stochastic partial differential equation (SPDE) on a Riemannian manifold. This approach allows not only to tackle the problem of defining GRFs on arbitrary compact smooth surfaces, but also the problem of accounting for transport phenomena in data lying on such surfaces. We also describe how prediction and inference based on such models relies on the manipulation of sparse matrices, thus opening the way to scalable computations.

2 The SPDE approach

As foretold by its name, the SPDE approach consists in using a stochastic partial differential equations (SPDE) to model the dynamic evolution of a GRF in time and space. This approach to GRF modeling has been popularized by Lindgren et al. (2011), and builds on a result from

Whittle (1954) which states that isotropic GRFs Z on $\mathbb{R}^d$ ($d \in \mathbb{N}$) with a Matérn covariance function are stationary solutions of the SPDE given by

$$(\kappa^2 - \Delta)^{\alpha/2} Z = \tau \mathcal{W}, \tag{1}$$

where $\kappa > 0$, $\alpha > d/2$, $\tau > 0$, and $(\kappa^2 - \Delta)^{\alpha/2}$ is a pseudo-differential operator (defined as $(\kappa^2 - \Delta)^{\alpha/2}[\cdot] = \mathcal{F}^{-1}[\boldsymbol{w} \mapsto (\kappa^2 + \|\boldsymbol{w}\|^2)^{\alpha/2} \mathcal{F}[\cdot](\boldsymbol{w})]$), and $\mathcal{W}$ is a Gaussian white noise on $\mathbb{R}^d$. Solving numerically this SPDE using stochastic finite elements allows to directly obtain an expression for the precision matrix (i.e. the inverse of the covariance matrix) of a GRF with Matérn covariance. Then, at the price of a minor approximation (called mass lumping), these expressions yield a Gaussian Markov random field representation of the GRF, characterized by a sparse precision matrix Lindgren et al. (2022); Rue and Held (2005). This in turn results in significant computational gains since sparse matrix algorithms can be used to deal with the matrix factorizations and linear system solving involved when performing sampling, prediction and inference Davis (2006); Lindgren et al. (2011).

The SPDE approach has been extensively used to model spatial data on Euclidean domains (see Lindgren et al. (2022) for a recent review), and extended to model spatial data on surfaces (see eg. Bonito et al. (2022); Coveney et al. (2019); Mejia et al. (2020)) and more generally on Riemannian manifolds (see eg. Herrmann et al. (2020); Lang and Pereira (2021)) by replacing the Laplace operator $-\Delta$ in SPDE (1) by a Laplace–Beltrami operator.

Extensions of the SPDE approach to the spatio-temporal setting have also been proposed. Cameletti et al. (2013) propose an approach where the spatial SPDE is coupled with an AR(1) process in time, thus yielding a separable model. Non-separable models based on a direct generalization of SPDE (1) have been proposed by Bakka et al. (2020) and Rayner et al. (2020), who consider the solutions of diffusion SPDE defined on Euclidean domains and on surfaces by

$$\frac{\partial Z}{\partial t} + (\kappa^2 - \Delta)^{\alpha/2} Z = \tau \mathcal{W}_T \otimes \mathcal{W}_S,$$

where $\mathcal{W}_T$ denotes a temporal white noise and $\mathcal{W}_S$ denotes either a white or colored noise in space. But these models result in random fields with even covariance functions, meaning that changing the sign of the spatial or temporal lag at which the covariance is evaluated does not change the value of the covariance. Consequently, these models are incapable of accounting for transport effects such as advection phenomena (which are intrinsically asymmetrical in time). Note however that in the Euclidean setting, extensions of the SPDE approach allowing to deal with asymmetries in the covariance structure have been proposed by Clarotto et al. (2022); Liu et al. (2022); Sigrist et al. (2015). However, to the best of our knowledge, the generalization of such models to more complex geometries is left open.

3 Advection-diffusion SPDEs on Riemannian manifolds

In their work, Clarotto et al. (2022) propose an extension of the SPDE approach to an advection-diffusion SPDE defined on a Euclidean domain, and derive analytically the expression of the covariance and precision matrices needed to perform sampling, prediction

and inference (through likelihood maximization).In this work, we propose to generalize their approach to the Riemannian setting by following the framework we proposed in earlier work Lang and Pereira (2021) to define spatial random fields on Riemannian manifolds.

More precisely, given a compact Riemannian manifold $(\mathcal{M}, g)$ defined from a smooth compact surface $\mathcal{M}$ equipped with a Riemannian metric g, we now consider the SPDE defined by We consider the following advection-diffusion SPDE on the domain $[0, T] \times \mathcal{M}$:

$$\frac{\partial \mathcal{Z}}{\partial t} + \frac{1}{c}\big(P(-\Delta_{\mathcal{M}})\mathcal{Z} + \operatorname{div}(\mathcal{Z}\gamma)\big) = \frac{\tau}{\sqrt{c}}\mathcal{W}_T \otimes \mathcal{Y}_S, \tag{2}$$

where P is a polynomial such for all $\lambda \geq 0$, $P(\lambda) \geq \kappa^2$, for some fixed $\kappa > 0$, $s \in \mathcal{M} \mapsto \gamma(s)$ is a smooth vector field on $T\mathcal{M}$ (the tangent bundle of $\mathcal{M}$), such that

$$\kappa^2 + \frac{1}{2}\operatorname{div}(\gamma) \geq \eta_0$$

for some fixed $\eta_0 > 0$, $\mathcal{W}_T \otimes \mathcal{Y}_S$ is a space-time separable stochastic forcing given as the product of a time-dependent Gaussian white noise $\mathcal{W}_T$ and a space-dependent colored noise $\mathcal{Y}_S = f_S(-\Delta_{\mathcal{M}})\mathcal{W}_S$ where $f_S : \mathbb{R}_+ \to \mathbb{R}$ is a bounded function, $c > 0$ is a time-scaling parameter, and $\tau > 0$ is a variance-scaling parameter. Since our goal is to work with arbitrary (smooth,compact) surfaces, we further assume that the advection vector field γ is given as the gradient of a scalar and differentiable function $\phi : \mathcal{M} \to \mathbb{R}$, i.e. $\gamma = \nabla_{\mathcal{M}}\phi$, which eliminates (for now) the question of how to properly define and parametrize vector fields on arbitrary (smooth, compact) surfaces.

We then formulate a space-time discretization scheme for SPDE (2) by combining the implicit Euler scheme proposed by Clarotto et al. (2022) with the Galerkin–Chebyshev approximation based on Surface finite elements (SFEM) of Lang and Pereira (2021). On the one hand, let $\mathcal{M}_h$ be a discretization of the manifold $\mathcal{M}$ into a polyhedral surface with mesh size $h > 0$ (by triangulation) which lies "close" to the surface $\mathcal{M}$. Let then $\{\psi_1, \ldots, \psi_N\} \subset H^1(\mathcal{M}_h)$ be the *linear* finite element basis associated with $\mathcal{M}_h$, where N is the number of nodes of the triangulation, and take $V_N = \operatorname{span}\{\psi_k : 1 \leq k \leq N\}$. We then denote by $\boldsymbol{C}$, $\boldsymbol{R}$ and $\boldsymbol{B}$ the matrices whose entries are respectively given by

$$C_{ij} = \langle \psi_i, \psi_j \rangle, \quad R_{ij} = \langle \nabla\psi_i, \nabla\psi_j \rangle, \quad B_{ij} = \langle \psi_i, \operatorname{div}(\gamma\psi_j) \rangle, \quad 1 \leq i, j \leq N,$$

and let $\sqrt{\boldsymbol{C}} \in \mathbb{R}^{N\times N}$ such that $\boldsymbol{C} = \sqrt{\boldsymbol{C}}(\sqrt{\boldsymbol{C}})^T$. We also introduce the scaled matrices $\tilde{\boldsymbol{R}} = (\sqrt{\boldsymbol{C}})^{-1}\boldsymbol{R}(\sqrt{\boldsymbol{C}})^{-T}$ and $\tilde{\boldsymbol{B}} = (\sqrt{\boldsymbol{C}})^{-1}\boldsymbol{B}(\sqrt{\boldsymbol{C}})^{-T}$. Note that all these matrices are sparse, due to the fact the functions ψ_i have small support which mostly do not intersect one another.

Finally, for some $K \in \mathbb{N}$, we write $\delta t = T/K$ and $t_k = k\delta t$ and introduce the matrix Γ defined by

$$\boldsymbol{\Gamma} = \boldsymbol{I} + \frac{\delta t}{c}\big(P(\widetilde{\boldsymbol{R}}) + \widetilde{\boldsymbol{B}}\big).$$

Then, the space-time discretization of SPDE (2) is the sequence $\{Z^{(k)}\}_{0\leq k\leq K}$ of functions of V_N resulting from the recursion:

$$\boldsymbol{z}^{(k+1)} = (\sqrt{\boldsymbol{C}})^{-T}\boldsymbol{\Gamma}^{-1}\left((\sqrt{\boldsymbol{C}})^T\boldsymbol{z}^{(k)} + \tau\sqrt{\frac{\delta t}{c}}f_S(\widetilde{\boldsymbol{R}})\boldsymbol{w}^{(k+1)}\right)$$

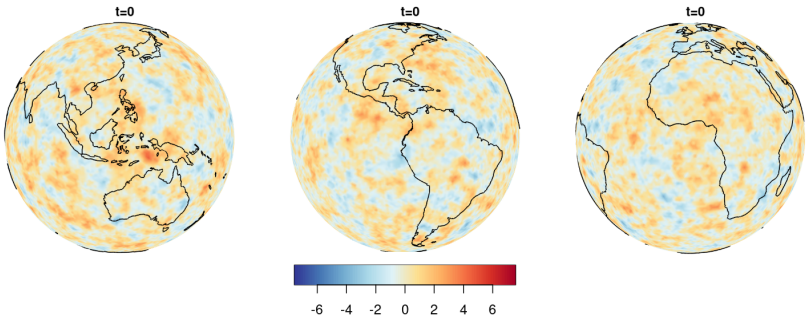


Figure 3.1: Simulation of a solution to the advection-diffusion SPDE on a meshed sphere, represented from three different viewpoints on the surface.

where we write $Z^{(k)} = \sum_{j=1}^{N} z_j^{(k)} \psi_j$.

We present in Figure 3.1 an example of simulation of SPDE (2) on a meshed sphere, resulting from the recursion above. We took P to be a polynomial of degree 1, and f_S to be the inverse of a polynomial of degree 1. The vector field was parametrized thanks to two scalar functions. As seen in the simulation, the resulting random fields seems to be able to recreate complex convection and diffusion phenomena.

This recursion can in turn be used to derive explicit formulas for the precision matrix $\boldsymbol{Q}$ of the vectors $\{\boldsymbol{z}^{(k)}\}_{0 \leq k \leq K}$ which in turn will only depend on the sparse matrices $\sqrt{\boldsymbol{C}}$, $\boldsymbol{\Gamma}$ and $\widetilde{\boldsymbol{R}}$. These vectors being Gaussian, explicit formula for their likelihood and conditional distribution are readily available and can be expressed using $\boldsymbol{Q}$. These formula are explicited in (Pereira; 2023).

Bibliographie

Bakka, H., Krainski, E., Bolin, D., Rue, H. and Lindgren, F. (2020). The diffusion-based extension of the Matérn field to space-time.

Bonito, A., Guignard, D. and Lei, W. (2022). Numerical approximation of Gaussian random fields on closed surfaces, *arXiv preprint arXiv:2211.13739* .

Cameletti, M., Lindgren, F., Simpson, D. and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach, *AStA Advances in Statistical Analysis* **97**: 109–131.

Chiles, J.-P. and Delfiner, P. (2009). *Geostatistics: Modeling Spatial Uncertainty*, Vol. 497, John Wiley & Sons.

Clarotto, L., Allard, D., Romary, T. and Desassis, N. (2022). The SPDE approach for spatio-temporal datasets with advection and diffusion, *arXiv preprint arXiv:2208.14015* .

Coveney, S., Corrado, C., Roney, C. H., Wilkinson, R. D., Oakley, J. E., Lindgren, F., Williams, S. E., O'Neill, M. D., Niederer, S. A. and Clayton, R. H. (2019). Probabilistic interpolation of uncertain local activation times on human atrial manifolds, *IEEE Transactions on Biomedical Engineering* **67**(1): 99–109.

Davis, T. A. (2006). *Direct Methods for Sparse Linear Systems*, Vol. 2, SIAM.

Herrmann, L., Kirchner, K. and Schwab, C. (2020). Multilevel approximation of Gaussian random fields: fast simulation, *Mathematical Models and Methods in Applied Sciences* **30**(01): 181–223.

Lang, A. and Pereira, M. (2021). Galerkin–Chebyshev approximation of Gaussian random fields on compact Riemannian manifolds, *arXiv preprint arXiv:2107.02667* .

Lindgren, F., Bolin, D. and Rue, H. (2022). The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running, *Spatial Statistics* **50**: 100599.

Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian markov random fields: the stochastic partial differential equation approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(4): 423–498.

Liu, X., Yeo, K. and Lu, S. (2022). Statistical modeling for spatio-temporal data from stochastic convection-diffusion processes, *Journal of the American Statistical Association* **117**(539): 1482–1499.

Mejia, A. F., Yue, Y., Bolin, D., Lindgren, F. and Lindquist, M. A. (2020). A Bayesian general linear modeling approach to cortical surface fMRI data analysis, *Journal of the American Statistical Association* **115**(530): 501–520.

Pereira, M. (2023). A note on spatio-temporal random fields on meshed surfaces defined from advection-diffusion SPDEs. working paper or preprint.
URL: *https://hal.science/hal-04132148*

Rayner, N. A., Auchmann, R., Bessembinder, J., Brönnimann, S., Brugnara, Y., Capponi, F., Carrea, L., Dodd, E. M., Ghent, D., Good, E. et al. (2020). The EUSTACE project: delivering global, daily information on surface air temperature, *Bulletin of the American Meteorological Society* **101**(11): E1924–E1947.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, CRC press.

Sigrist, F., Künsch, H. R. and Stahel, W. A. (2015). Stochastic partial differential equation based modelling of large space-time data sets, *Journal of the Royal Statistical Society: Series B: Statistical Methodology* pp. 3–33.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*, Springer Science & Business Media.

Whittle, P. (1954). On stationary processes in the plane, *Biometrika* pp. 434–449.

Deep kernel learning for geostatistics

Thomas Romary [1], Solal Raymondjean [1] & Nicolas Desassis [1]

[1] *Equipe Géostatistique, Centre de Géosciences, Mines Paris PSL, Fontainebleau, 77300, France, thomas.romary@minesparis.psl.eu*
[2] *Equipe Géostatistique, Centre de Géosciences,, Mines Paris PSL, Fontainebleau, 77300, France, nicolas.desassis@minesparis.psl.eu*

Résumé. Les processus gaussiens représentent une riche classe de modèles pour l'apprentissage supervisé, en particulier pour la cartographie des variables naturelles en géostatistique. Ils sont particulièrement intéressants dans la mesure où ils permettent de quantifier la précision de la prédiction en de nouveaux points de données. Les implémentations traditionnelles reposent sur l'hypothèse de stationnarité du second ordre. Cette hypothèse n'est généralement pas respectée lorsque l'on considère de grands jeux de données provenant de réseaux de capteurs modernes ou d'images satellite. Diverses constructions ont été proposées pour surmonter cette limitation, telles que les approches à paramètres variables dans les modèles de convolution ou dans les équations différentielles partielles stochastiques, mais elles restent difficiles à implémenter. Dans ce travail, nous considérons l'approche par déformation de l'espace où l'estimation de la déformation est vue comme un problème de transport. À cet égard, nous modélisons la fonction de déformation en nous inspirants des flots normalisés, qui sont particulièrement adaptés à ce problème. Les flots normalisés apprennent classiquement à transformer une distribution donnée, généralement facile à échantillonner, en une distribution de données. Ici, nous considérons plutôt l'apprentissage d'une transformation bijective de l'espace géographique d'entrée à l'espace déformé, où la stationnarité et l'isotropie sont vérifiées. La procédure d'estimation est basée sur l'optimisation par descente de gradient stochastique sur la log-vraisemblance négative, qui est mise à l'échelle en utilisant l'approximation de Vecchia. L'approche est illustrée à la fois sur des données synthétiques et sur un jeu de données réelles.

Mots-clés. Processus gaussiens, non-stationnaire, apprentissage profond, approximation de Vecchia

Abstract. Gaussian processes represent a rich class of models for supervised learning, particularly for mapping natural variables in geostatistics. They are especially appealing in that they provide a quantification of the prediction accuracy at new data points. Traditional implementations involve second order stationarity hypothesis. This assumption is generally violated when considering large datasets issued from modern sensor networks or satellite images. Various constructions have been proposed to overcome this limitation, such as varying parameter approaches in convolution models or in stochastic partial differential equations, but they remain difficult to handle. In this work, we consider the space deformation approach where the estimation of the deformation is considered as a transport problem. In this regard, we model the warping function with neural networks inspired by normalizing flows, which are particularly taylored to this problem. Normalizing flows classically learn to

transform a given, generally easy to sample, distribution into the data distribution. Here, we consider instead learning a deep bijective transformation from the input geographic space to the warped space, where stationarity and isotropy hold. The estimation procedure is based on stochastic gradient descent optimization over the negative loglikelihood, which is scaled up using the Vecchia approximation. The approach is illustrated both on synthetic data and a real dataset.

Keywords. Gaussian process, non-stationary, deep learning, Vecchia approximation

1 Introduction

Les processus gaussiens sont couramment employés en géostatistique pour modéliser des variables naturelles ou bien de manière latente dans la modélisation d'événements ponctuels Gelfand et al. (2011). L'hypothèse de stationnarité de la fonction de covariance qui est généralement faite est parfois difficilement tenable lorsqu'on considère de grands ensembles de données, pouvant notamment faire apparaître des anisotropies variables localement. Différentes approches ont été proposées dans la littérature pour proposer des constructions non-stationnaires de la fonction de covariance, notamment l'approche par déformation de l'espace (Sampson and Guttorp, 1992; Fouedjio et al., 2015), par convolution (Paciorek and Schervish, 2006; Fouedjio et al., 2016) ou bien en faisant varier localement l'opérateur de diffusion dans les équations aux dérivées partielles stochastiques Lindgren et al. (2011); Fuglstad et al. (2015); Pereira (2019). Leurs implémentations sont cependant complexes à mettre en œuvre et l'estimation des paramètres par maximum de vraisemblance est rendue difficile dans un contexte de grands jeux de données. Dans ce cadre, l'approximation de Vecchia (Vecchia, 1988; Katzfuss and Guinness, 2021) est particulièrement performante et permet de limiter la complexité des problèmes d'estimation des paramètres et de prédiction, qui font tous deux intervenir l'inversion de la matrice de covariance des données.

De précédentes tentatives de tirer parti des réseaux de neurones profonds pour des problèmes de prédiction de champs aléatoires en des emplacements non observées (Chen et al., 2020; Zammit-Mangion et al., 2021) reposent sur une projection préalable des données dans une base de fonctions spatiales, ce qui à tendance à lisser les dépendances locales.

Dans ce travail, nous proposons d'adapter l'approche de Wilson et al. (2016) à la problématique de la caractérisation d'une fonction de déformation de l'espace. Ce problème peut être vu comme un problème de transport de la localisation des points de données vers un espace où les hypothèses de stationnarité et d'isotropie sont vérifiées. Pour le résoudre, nous proposons des architectures de deep learning adaptées, inspirées des normalizing flows (Papamakarios et al., 2021), pour modéliser la fonction de déformation. L'apprentissage est réalisé par maximum de vraisemblance, en tirant parti de l'approximation de Vecchia pour accélérer les calculs.

2 Processus Gaussiens

On considère un jeu de données $\mathrm{Z} = (Z(x_1), \ldots, Z(x_n)) = (Z_1, \ldots, Z_n) \in \mathbb{R}^n$, où $x_i \in \mathbb{R}^d$. On fait l'hypothèse que Z est tiré d'une réalisation d'un processus gaussien Z, d'espérance μ et de fonction de covariance K_θ, paramétrée par θ. Par conséquent,

$$\mathrm{Z} \sim \mathcal{N}(\mu, \mathrm{K}),$$

où $\mathrm{K}_{i,j} = K_\theta(Z(x_i), Z(x_j))$. Dans ce cadre, si $x_0 \in \mathbb{R}^d$, on a

$$Z(x_0)|\mathrm{Z} \sim \mathcal{N}(\mu + k_0^t \mathrm{K}^{-1}(\mathrm{Z} - \mu), \sigma_0^2 - k_0^t \mathrm{K}^{-1} k_0),$$

où $(k_0)_i = K_\theta(x_0, x_i)$ et $\sigma_0^2 = K_\theta(x_0, x_0)$.

3 Déformation de l'espace

L'approche de modélisation par déformation de l'espace consiste à chercher un difféomorphisme f, tel que $\mathrm{Cov}(Z(f(x_1)), Z(f(x_2))) = K'(\|x_1 - x_2\|)$, où K' est une fonction de covariance stationnaire isotrope. On peut voir en figure 1 un exemple d'une telle transformation. f résoud ainsi un problème de transport du plan d'échantillonnage initial vers un espace où les données se comportent de façon stationnaire. Si l'on considère que le plan d'échantillonnage $X = (x_1, \ldots, x_n)$ est tiré selon une densité p_x, alors la densité p_f correspondant au plan d'échantillonnage transformé vérifie :

$$p_f(u) = p_x(f^{-1}(u))|\mathrm{det} J_{f^{-1}}(u)|,$$

où $J_{f^{-1}}$ est le jacobien de f^{-1}. Cette relation suggère de modéliser f à l'aide de normalizing flows, qui consistent en une succession de transformations bijectives faisant intervenir des réseaux de neurones simples, et que nous noterons f_θ.

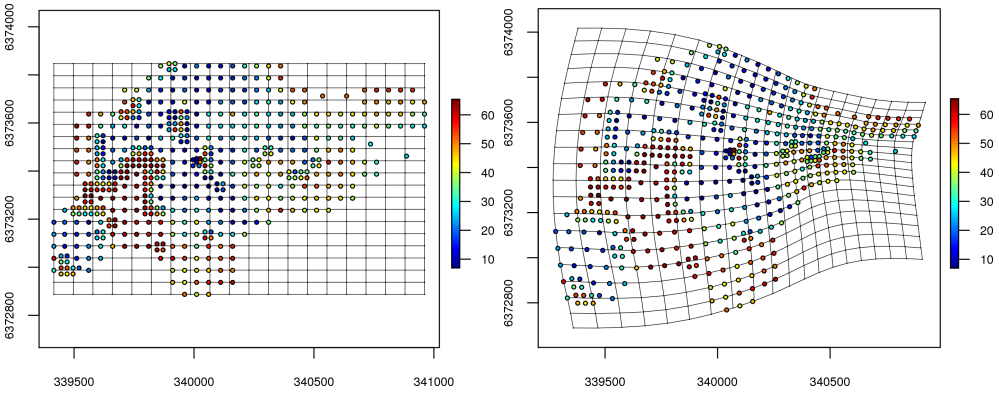


Figure 1: Exemple de déformation de l'espace: espace géographique à gauche, espace déformé à droite. Source : Fouedjio et al. (2015).

Finalement, l'estimation des paramètres du modèle de processus gaussien non-stationnaire repose sur la maximisation de la vraisemblance jointe de Z et du plan d'échantillonnage X, qui s'écrit :

$$\begin{aligned} p(Z, X) &= p(Z|X)p(f_\theta(X)) \\ &= \mathcal{N}(Z; \mu, \mathrm{K}_\theta)p_x(X)|\mathrm{det}J_{f_\theta}(X)|^{-1}, \end{aligned}$$

où $(\mathrm{K}_\theta)_{i,j} = K(f_\theta(x_i), f_\theta(x_j))$.

4 Approximation de Vecchia

Le principe de l'approximation de Vecchia repose sur l'observation suivante :

$$p(\mathrm{Z}) = p(Z_1)\prod_{i=2}^{n} p(Z_i|Z_{<i}),$$

où $\{< i\}$ désigne l'ensemble des indices inférieurs à i. En choisissant avec soin des sous-ensembles d'indices $c(i)$ parmi $\{< i\}$ de taille au plus $m << n$, on obtient l'approximation

$$p(\mathrm{Z}) \approx p(Z_1)\prod_{i=2}^{n} p(Z_i|Z_{c(i)}).$$

Pour un ensemble de $c(i)$ donnés, cette approximation est optimale au sens de la divergence de Kullback-Leibler (Schäfer et al., 2021). Dans le cas des processus gaussiens, elle permet d'obtenir une approximation de la matrice de covariance des données sous la forme $\mathrm{K}^{-1} \approx UU'$, où U est une matrice triangulaire supérieure, dont les éléments sont donnés par :

$$U_{j,i} = \begin{cases} \left(\sigma_i^2 - k_i \mathrm{K}_{c(i)}^{-1} C_i^t\right)^{-1/2} & \text{if } i = j \\ -(\mathrm{K}_{c(i)}^{-1} k_i)_j U_{i,i} & \text{if } j \in c(i) \\ 0 & \text{otherwise} \end{cases}$$

où $k_i = K(Z_i, c(i))$ et $\mathrm{K}_{c(i)} = K(Z_{c(i)}, Z_{c(i)})$, et $U_{j,i} = -(\mathrm{K}_{c(i)}^{-1} k_i)_j U_{i,i}$ si $j \in c(i)$, 0 sinon. Chaque ligne contient donc au plus $m+1$ éléments non nuls. Dans un problème de prédiction, une approximation similaire est construite en considérant la matrice de covariance du vecteur des cibles et des données conjointement.

5 Conclusion

Notre approche combine l'estimation de la fonction de déformation par réseaux de neurones profonds avec l'approximation de Vecchia. A noter que l'approche nécessite de mettre à jour les ensembles $c(i)$ à chaque itération de l'optimisation afin de conserver une approximation précise. Différentes architectures pour la modélisation de f sont explorées et comparées avec des approches standards, sur des données synthétiques et un jeu de données réelles.

Bibliographie

Chen, W., Li, Y., Reich, B. J., and Sun, Y. (2020). Deepkriging: Spatially dependent deep neural networks for spatial prediction. *arXiv preprint arXiv:2007.11972*.

Fouedjio, F., Desassis, N., and Rivoirard, J. (2016). A generalized convolution model and estimation for non-stationary random functions. *Spatial Statistics*, 16:35–52.

Fouedjio, F., Desassis, N., and Romary, T. (2015). Estimation of space deformation model for non-stationary random functions. *Spatial Statistics*, 13:45 – 61.

Fuglstad, G.-A., Lindgren, F., Simpson, D., and Rue, H. (2015). Exploring a new class of non-stationary spatial gaussian random fields with varying local anisotropy. *Statistica Sinica*, pages 115–133.

Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P., editors (2011). *Handbook of Spatial Statistics*. Chapman & Hall/CRC.

Katzfuss, M. and Guinness, J. (2021). A general framework for vecchia approximations of gaussian processes. *Statistical Science*, 36(1):124–141.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B*, 73(4):423–498.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64.

Pereira, M. (2019). *Champs aléatoires généralisés définis sur des variétés riemanniennes: théorie et pratique.* PhD thesis, Mines ParisTech.

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119.

Schäfer, F., Katzfuss, M., and Owhadi, H. (2021). Sparse cholesky factorization by kullback–leibler minimization. *SIAM Journal on scientific computing*, 43(3):A2019–A2046.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):297–312.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. (2016). Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378. PMLR.

Zammit-Mangion, A., Ng, T. L. J., Vu, Q., and Filippone, M. (2021). Deep compositional spatial models. *Journal of the American Statistical Association*, 0(0):1–22.

Traitement automatique du langage naturel (Groupe MALIA)

SigBERT: Combiner des compte-rendus médicaux de patients et la théorie des Signatures de chemins pour la Prédiction de Survie

Paul Minchella [1] & Loic Verlingue [2] & Stéphane Chrétien [3] & Rémi Vaucher [3] & Guillaume Metzler [3]

[1] *SHAPE-Med@Lyon, France, paul.minchella@lyon.unicancer.fr*
[2] *Centre Léon Bérard, France, loic.verlingue@lyon.unicancer.fr*
[3] *Laboratoire ERIC, France, {stephane.chretien, r.vaucher, guillaume.metzler, }@univ-lyon2.fr*

Résumé. Les rapports médicaux électroniques contiennent une quantité considérable d'informations exploitables pour les applications d'apprentissage automatique en santé. Cependant, les méthodes existantes d'analyse de survie peinent souvent à traiter efficacement la complexité et l'hétérogénéité de ces données, en particulier lorsqu'il s'agit de dossiers patients séquentiels. Nous proposons ici SigBERT, un pipeline novateur d'analyse de survie temporel en oncologie capable de gérer efficacement un grand nombre de documents cliniques textuels associés à chaque patient, afin d'améliorer l'estimation du risque.

SigBERT prend en entrée les comptes rendus médicaux pour chaque patient, tous datés. Il extrait et moyenne les word embeddings en un vecteur de sentence embedding pour chaque rapport. On applique l'extraction par signatures aux séries temporelles que forment les coordonnées de ces vecteurs afin d'en extraire des features. Ces coefficients de signatures, capturant les propriétés géométriques et séquentielles des trajectoires, sont enfin injectés dans un modèle de Cox, combiné au LASSO, pour estimer le score de risque d'un patient.

Le modèle a été entraîné sur un jeu de données oncologiques réel comprenant 3 560 patients et évalué sur une cohorte indépendante de 3 561 patients issue des corpus du Centre Léon Bérard. Sur l'ensemble de validation, il a atteint un C-index moyen de 0.75 (sd 0.014), un td-AUC de 0.800 (sd 0.022) avant 10 ans, et un Integrated Brier Score de 0.035 (sd 0.014) avant 3 ans, attestant de solides performances prédictives et discriminantes.

SigBERT exploite les données médicales séquentielles pour affiner l'estimation du risque, contribuant ainsi à l'avancement de l'analyse de survie basée sur les rapports narratifs.

Mots-clés. NLP, Extraction par Signature des chemins, Analyse de Survie, Oncologie

Abstract. Electronic medical reports contain a vast amount of information that can be leveraged for machine learning applications in healthcare. However, existing survival analysis methods often struggle to effectively handle the complexity and heterogeneity of sequential data. Here, we propose SigBERT, an innovative temporal survival analysis framework designed to efficiently process a large number of textual clinical reports per patient, thereby improving risk estimation.

SigBERT processes timestamped medical reports by extracting word embeddings, averaging them into sentence embeddings, and constructing a time series. To capture temporal

dynamics, we apply signature extraction from rough path theory, encoding both geometric and sequential patterns of patient trajectories. These features are then integrated into a LASSO-penalized Cox model to estimate patient-specific risk scores.

The model was trained on a real-world oncology dataset of 3,560 patients and evaluated on an independent cohort of 3,561 patients from the Centre Léon Bérard corpus. It achieved on validation set a mean C-index score of 0.75 (sd 0.014), a td-AUC of 0.800 (sd 0.022) before 10 years and an Integrated Brier Score of 0.035 (sd 0.014) before 3 years, which demonstrates solid predictive discriminated performance.

SigBERT integrates sequential medical data to enhance risk estimation, advancing narrative-based survival analysis.

Keywords. NLP, Signature coefficient Extraction, Survival Analysis, Oncology

1 Introduction

L'analyse de survie joue un rôle fondamental en médecine (oncologie, cardiologie, néphrologie, soins intensifs, ...), où la prédiction du pronostic des patients est essentielle pour orienter la prise de décision clinique. Elle permet de déterminer les stratégies de traitement, d'affiner les critères d'éligibilité aux essais cliniques, ou encore faciliter la stratification des risques. Le modèle de Cox avec risques proportionnels, présenté la première fois par Cox (1972), est depuis sa publication la référence majeure en analyse de survie en raison de son interprétabilité et de son efficacité dans l'identification des facteurs pronostiques. Cependant, ce modèle s'appuie uniquement sur un instant connu des variables, rendant ces dernières statiques au moment de l'inclusion des patients dans l'étude. Un cadre d'analyse de survie robuste devrait intégrer des données patients reflétant l'évolution temporelle de leur suivi, afin d'améliorer la précision des prédictions.

2 Cohorte

Cette étude est conforme au Règlement Général sur la Protection des Données (RGPD) et s'inscrit dans le cadre de la recherche scientifique relevant de l'intérêt légitime de la recherche en cancérologie, conformément aux articles 6.1.f et 9.2.j du Règlement (UE) n° 2016/679. Ce projet a été officiellement enregistré sous la déclaration MR004 (V3.2, 23/08/2021) au Centre Léon Bérard, garantissant le respect des normes légales et éthiques en matière de traitement des données de santé. Les données ont été soigneusement anonymisées et ne peuvent être utilisées que dans le cadre de cette étude. Aucun patient ne s'est opposé à l'étude.

Le jeu de données est composé de corpus textuels propres, contenant 274 420 rapports médicaux issus de 7 121 patients, dont 4 983 sont décédés et 2 138 sont censurés. Les rapports sont principalement des comptes rendus de consultation (63%) et des comptes rendus de

séjour hospitalier (32%). Chaque patient possède en moyenne 39 rapports médicaux (écart-type: 25). Ce jeu de données couvre tous les types de cancers, dont les plus représentés sont les cancers du sein (25%), gynécologique (9.7%), intestinaux (8.6%), pulmonaire (5.5%), prostate (7%) et endocrinien (4.6%). 90% des patients ont été diagnostiqués entre avril 1997 et septembre 2020. La durée médiane d'étude est de 1 024 jours, avec 75% des événements survenant avant 2 093 jours.

3 Méthode

Nous visons donc à développer un pipeline général qui prend en entrée l'ensemble des compte rendus médicaux datés pour chaque patient, pour montrer l'impact des données textuelles dans l'estimation de la survie. Pour éviter de biaiser les résultats, on masque les compte-rendus des (au moins) 100 derniers jours précédant le décès, s'il a lieu. Les sections qui suivent détaillent le pipeline présenté en Figure 1.

3.1 Extraction d'embeddings

Chaque rapport médical est vectorisé via l'extraction de word embeddings, encodant ainsi l'information sémantique de chaque token dans un espace de dimension $p = 768$. Pour ce faire, on applique OncoBERT par Vienne et al. (2024), un modèle de langage affiné sur un large corpus de notes cliniques en oncologie du Centre Léon Bérard. Pour un patient i à l'instant t, cela s'écrit :

$$\{\text{word}_1 : v_{w_1},\ \ldots,\ \text{word}_{N_t} : v_{w_{N_t}}\}_i \tag{1}$$

où N_t représente le nombre de mots uniques (tokens) présents dans le rapport médical à l'instant t. On moyenne ensuite ces représentations v_w en un vecteur unique pour chaque rapport médical, grâce à la robuste méthode non supervisée *Smooth Inverse Frequency* (SIF) par Arora et al. (2017). Pour un rapport s, sa représentation est donnée par :

$$v_s = \frac{1}{|s|} \sum_{w \in s} \frac{a}{\mathrm{f}(w) + a}\ v_w\ \in \mathbb{R}^p \tag{2}$$

où $\mathrm{f}(w)$ représente la fréquence d'apparition d'un mot donné dans le corpus, et a est un paramètre de lissage (généralement fixé à 10^{-3}). Ainsi, on peut formellement représenter un patient avec N_i rapports par une famille de p séries temporelles par :

$$\left(v_{t_1}, \ldots, v_{t_{N_i}}\right)_i^\top \in \mathbb{R}^{N_i \times p} \tag{3}$$

Motivé par le lemme de Johnson and Lindenstrauss (1984), et pour éviter un nombre trop grand de coefficients à calculer, on peut compresser chaque vecteur $v_s \in \mathbb{R}^p$ en multipliant ce dernier par une matrice $R_{\text{comp}} \in \mathbb{R}^{\bar{p} \times p}$, tant que les entrées de cette dernière sont i.i.d, centrées et de variance normalisée par $1/\bar{p}$. On considèrera $v_s \leftarrow R_{\text{comp}} \cdot v_s$. L'information sémantique ne s'en retrouve que très peu impactée, mais cela permet de réduire la dimension des chemins de $p = 768$ à $\bar{p} = 25$, par exemple.

3.2 Extraction de covariables par signature

La signature d'un chemin, issue de la théorie des chemins rugueux par Chen (1954), fournit une méthode systématique pour encoder des données séquentielles comme en (3), à l'aide d'intégrales itérées. Considérons donc un chemin de dimension p, noté $v = (v^1, \dots, v^p)$, défini sur l'intervalle $[0, T]$. Pour tout entier $k \geq 1$, pour tout $0 \leq t_1 < \cdots < t_k \leq T$ et toute suite d'indices $i_1, \dots, i_k \in \{1, \dots, p\}$, la signature d'intégrale itérée de v jusqu'au temps $t \in [0, T]$ est définie par :

$$S(v)_{0,t}^{(i_1,\dots,i_k)} = \int_{0<t_k<t} \cdots \int_{0<t_1<t_2} \mathrm{d}v_{t_1}^{i_1} \dots \mathrm{d}v_{t_k}^{i_k} \tag{4}$$

En plus d'encoder la dynamique temporelle, cette approche gère des séquences de longueurs variables, est invariante par translation et reparamétrisation temporelle (voir Chevyrev and Kormilitzin (2016)). De plus, l'identité de Chen optimise le calcul en exprimant tout chemin concaténé comme le produit des signatures de ses sous-chemins. Pour un patient i, l'ensemble des covariables extraites sur sa plage temporelle de suivi, tronqué à l'ordre 3, est noté :

$$\mathbb{S}_i = \left(S^{(1)}, \dots, S^{(p)}, S^{(1,1)}, \dots, S^{(p,p)}, S^{(1,1,1)}, \dots, S^{(p,p,p)} \right)_i \tag{5}$$

3.3 Modèle de survie

Notre choix pour illustrer l'impact des données textuelles s'est tourné vers le modèle de Cox (1972), qui prend en compte les patients censuré, c'est-à-dire ceux dont l'événement d'intérêt $\mathcal{T}$ (décès, rechute, etc.) n'a pas encore eu lieu. Ces observations contribuent elles-aussi à l'estimation de la vraisemblance, permettant ainsi de limiter les biais et d'améliorer la robustesse des prédictions. L'objectif est d'estimer la probabilité de survie d'un patient jusqu'à l'instant t, conditionnellement à ses covariables $\mathbb{S}$, soit $\mathcal{S}(t \mid \mathbb{S}) := \mathbb{P}\big(\mathcal{T} \geq t \mid \mathbb{S}\big)$. Cette estimation repose sur la notion clé de risque instantané de décès h, qui mesure la probabilité infinitésimale de survenue de l'événement à t et se relie à la survie par l'équation suivante :

$$\mathcal{S}(t \mid \mathbb{S}) = \exp\left(- \int^t h(s \mid \mathbb{S}) \, \mathrm{d}s \right) \tag{6}$$

Cox propose donc le modèle linéaire généralisé :

$$h(t \mid \mathbb{S}) \; = \; h_0(t) \cdot \exp\big(\mathbb{S} \cdot \beta\big) \tag{7}$$

où $\beta \in \mathbb{R}^d$ est le vecteur de paramètres à estimer et h_0 est le risque de base, commun à tous les patients, estimé par Breslow (1972). On pose $\eta := \mathbb{S} \cdot \beta$, appelé score de risque. Pour estimer β, on doit traiter un nombre de covariables naturellement élevé. Puisque ce déséquilibre accroît le risque de surapprentissage, on propose l'emploi de la méthode LASSO (*Least Absolute Shrinkage and Selection Operator*) appliqué au modèle de Cox, d'après Tibshirani (1997) :

$$\widehat{\beta}^{\ell_1} \in \underset{\beta}{\operatorname{argmax}} \; \log \mathrm{PL}(\beta) - \lambda^{(1)} \|\beta\|_1 \tag{8}$$

où $\mathrm{PL}(\beta)$ est la vraisemblance partielle par Cox (1975), et $\lambda^{(1)} > 0$ le paramètre de régularisation.

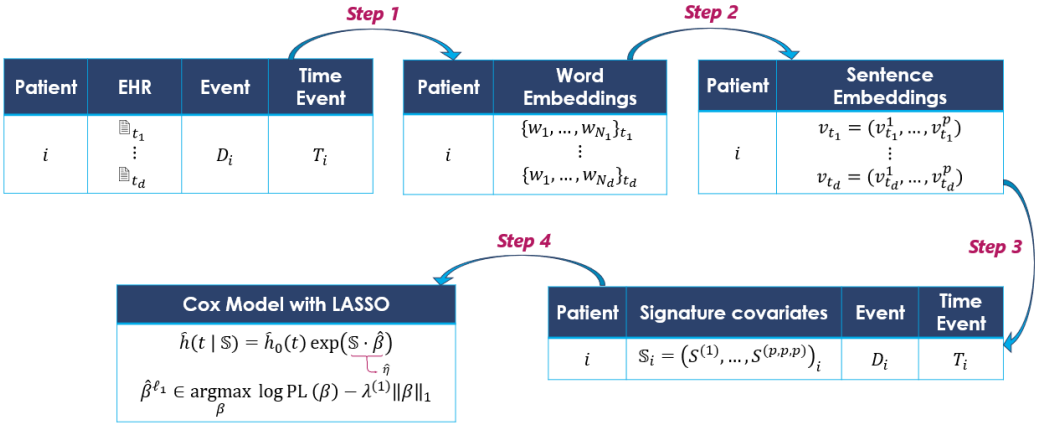


Figure 1: Résumé global de notre pipeline

4 Résultats

Le modèle a été entraîné sur 3 560 patients (136 748 rapports) et évalué sur un jeu de validation de 3 561 patients (137 672 rapports), divisés en 10 sous-ensembles indépendants. Pour évaluer la qualité du modèle, on invoque l'indice de concordance (C-index), métrique fondamentale en analyse de survie, définie en Harrell et al. (1982). Il mesure la proportion de paires concordantes parmi toutes les paires possibles. Si j connaît l'évènement avec i, alors le risque associé à j se doit d'être plus élevé.

$$\text{C-index} = \frac{\sum_{i,j} \mathbb{1}_{\{T_j<T_i\}} \cdot \mathbb{1}_{\{\widehat{\boldsymbol{\eta}}_j>\widehat{\boldsymbol{\eta}}_i\}} \cdot \delta_j}{\sum_{i,j} \mathbb{1}_{\{T_j<T_i\}} \cdot \delta_j} \tag{9}$$

où T_i, T_j sont les temps dans l'étude des individus i et j, $\widehat{\boldsymbol{\eta}}_i$, $\widehat{\boldsymbol{\eta}}_j$ sont les risques estimés respectifs et $\delta_j = 1$ si j a connu l'évènement, 0 sinon. Un C-index $= 1$ caractérise un modèle parfait, tandis qu'un C-index $= 0.5$ correspond à une performance équivalente au hasard. Un C-index supérieur à 0.7 est généralement considéré comme satisfaisant. Son penchant temporel, le td-AUC, est défini puis intégré sur l'intervalle temporel d'intérêt. L'erreur de calibration est quantifiée par le Score de Brier (BS), également moyenné sur la plage temporelle choisie. Un BS inférieur à 0.25 est considéré comme un bon indicateur de calibration. Les expressions analytiques de ces métriques sont détaillées dans Pölsterl (2019).

Notre modèle obtient un C-index moyen de 0.75 (sd 0.014) avec un $\text{IC}_{0.95}$: [0.7417, 0.7593], indiquant une bonne capacité discriminante. Le td-AUC moyen sur 10 ans est de 0.800 (sd 0.022), et le Score de Brier intégré reste inférieur à 0.25 jusqu'à 10 ans, avec 0.0347 (sd 0.0014) à 3 ans, témoignant d'une calibration satisfaisante.

Ces résultats s'alignent avec les performances des modèles État de l'Art, tout en étant obtenus dans un contexte clinique réel complexe, renforçant la crédibilité de notre modèle. De plus, l'emploi de la compression, de la méthode LASSO et l'identité de Chen des signatures permet un apprentissage et une inférence efficace, de l'ordre de quelques minutes une fois les word embeddings extraits, soulignant un autre aspect unique de notre modèle.

5 Discussion

Notre approche illustre le potentiel d'exploitation des données séquentielles textuelles pour l'analyse de survie en oncologie. Ces premiers résultats satisfaisants ouvrent la voie à une extension du modèle à l'ensemble des patients de la base de données, permettant une généralisation plus large de l'approche. Une amélioration possible réside dans l'intégration de données tabulaires et séquentielles, en complément des embeddings textuels. En combinant données narratives et structurées (caractéristiques du patient, de la maladie, données biologiques, ...), nous visons à construire un modèle de survie multimodal, capable d'offrir une prise en charge plus personnalisée et précise en oncologie.

Ce travail a été financièrement soutenu par le programme Shape-MED@Lyon.

Bibliographie

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations (ICLR)*. Published as a conference paper at ICLR 2017.

Breslow, N. E. (1972). Contribution to the discussion of the paper by d. r. cox. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):216–217.

Chen, K.-T. (1954). Iterated integrals and exponential homomorphisms. *Proceedings of the London Mathematical Society*, 4:502–512.

Chevyrev, I. and Kormilitzin, A. (2016). A primer on the signature method in machine learning.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of lipschitz mappings into a hilbert space. In *Contemporary Mathematics*, volume 26, pages 189–206. American Mathematical Society.

Pölsterl, S. (2019). Evaluating survival models. Accessed: 2025-01-14.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395.

Vienne, R., Filori, Q., Susplugas, V., Crochet, H., and Verlingue, L. (2024). Abstract 3475: Prediction of nausea or vomiting, and fatigue or malaise in cancer care. *Cancer Research*, 84:3475–3475.

Large Language Models as Markov Chains

Oussama Zekri [1*] Ambroise Odonnat [2*]
Abdelhakim Benechehab[3] Linus Bleistein[4] Nicolas Boullé[5] Ievgen Redko[6]

[1] *ENS Paris-Saclay, France, oussama.zekri@ens-paris-saclay.fr*
[2] *Noah's Ark Lab & Inria, France, ambroiseodonnattechnologie@gmail.com*
[3] *Noah's Ark Lab & Eurecom, France, abdelhakim.benechehab@gmail.com*
[4] *Inria, Univ. Paris-Saclay, UEVE, France, linus.bleistein@gmail.com*
[5] *Imperial College London, United-Kingdom, n.boulle@imperial.ac.uk*
[6] *Noah's Ark Lab, France, ievgen.redko@gmail.com*
[*] *Equal contribution*

Résumé. Les grands modèles de langage (LLM) sont remarquablement efficaces dans une large gamme d'applications de traitement du langage naturel et bien au-delà. Cependant, une analyse théorique complète des capacités de généralisation des LLM reste à faire. Dans notre article, nous abordons cette tâche en établissant une équivalence entre les modèles de langage basés sur des transformers autorégressifs et les chaînes de Markov définies sur un espace d'état fini. Cela nous permet d'étudier rigoureusement le mécanisme d'inférence à plusieurs étapes des LLM. Nous établissons un lien entre les résultats obtenus et les comportements observés avec les LLM, tels que les répétitions et les réponses incohérentes. Enfin, nous utilisons notre formalisation pour dériver des bornes de généralisation de pré-entraînement et d'apprentissage en contexte pour les LLMs sous des données et hypothèses réalistes. Des expériences avec les modèles récents tels que Llama et Gemma montrent que notre théorie prédit correctement leur comportement dans la pratique.

Mots-clés. Modelès de langage, chaînes de Markov, apprentissage en contexte, borne de généralisation, sample complexity

Abstract. Large language models (LLMs) are remarkably efficient across a wide range of natural language processing tasks and well beyond them. However, a comprehensive theoretical analysis of the LLMs' generalization capabilities remains elusive. In our paper, we approach this task by drawing an equivalence between autoregressive transformer-based language models and Markov chains defined on a finite state space. This allows us to study the multi-step inference mechanism of LLMs from first principles. We relate the obtained results to the pathological behavior observed with LLMs such as repetitions and incoherent replies with high temperature. Finally, we leverage the proposed formalization to derive pre-training and in-context learning generalization bounds for LLMs under realistic data and model assumptions. Experiments with the most recent Llama and Gemma herds of models show that our theory correctly captures their behavior in practice.

Keywords. Large language models, Markov chain, in-context learning, generalization bounds, sample complexity

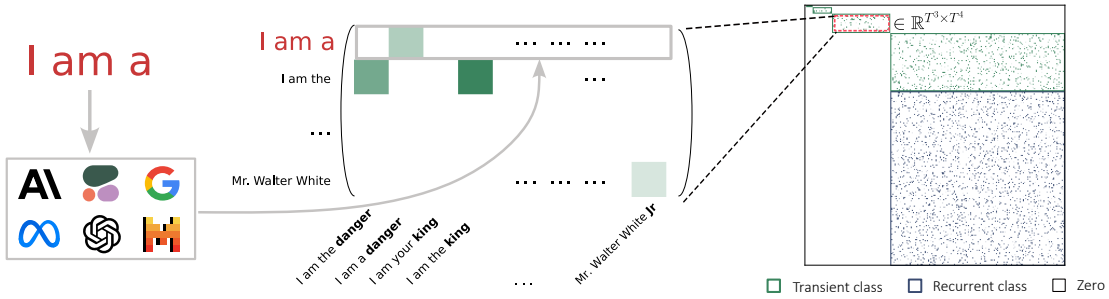


Figure 1: *LLM as a Markov chain.* A large language model with vocabulary size T and context window K is equivalent to a Markov chain with a sparse and block-structured transition matrix of size $\sum_{i \leq K} T^i \sim \mathcal{O}(T^K)$. The latter captures all possible outputs of a given LLM for all possible input sequences allowed by its vocabulary and context window.

1 Introduction

Large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023) have reshaped the landscape of machine learning and artificial intelligence. Trained on vast amounts of data, they have reached unprecedented performance in tasks such as machine translation (Brown et al., 2020), text generation, question answering (Roberts et al., 2020), and sentiment analysis (Zhang et al., 2023a), to name a few. Pre-trained LLMs also exhibit an intriguing capability of performing learning from the context, also called in-context learning (ICL), without explicitly updating their parameters.

Despite all these remarkable achievements, the theoretical justification for LLMs' impressive performance remains elusive. The two main reasons for this are 1) the complexity of analyzing modern deep transfer-based architectures used to train LLMs and 2) the non-iid nature of their pre-training data. Prior works usually bypass these challenges by imposing simplifying assumptions. Such assumptions include considering a shallow single-head transformer (Makkuva et al., 2024), only self-attention mechanism (Ildiz et al., 2024) or attention-only transformers (Edelman et al., 2024; Jeon et al., 2024) when studying LLMs. The others are to study the auto-regressive mechanism of the LLMs without taking into account their architecture explicitly (Lotfi et al., 2023, 2024; Xie et al., 2022). When the architecture is studied in full generality (Li et al., 2023; Zhang et al., 2023b), it is also common to relax the non-iid assumption by a certain predefined Markovian process.

This work provides a unified theoretical analysis of LLM's inference, pre-training, and ICL under realistic assumptions on model architecture and pre-training data. Our core idea is to note that despite the seemingly infinite generation capacity of LLMs, their vocabulary and context window are finite, making all possible input and output sequences countable. This leads to an intuitive, yet overlooked, approach that interprets LLMs as Markov chains whose transition matrix operates on a finite state space of such sequences (see Fig. 1). The generalization of LLMs is then their capacity to learn the transition probabilities of language itself.

Main contributions. Our contributions can be summarized as follows.

1) We characterize the inference mechanism of any LLM as a *finite-state* Markov chain. This provides a deeper understanding of LLMs' generation behavior and sheds new light on some of their pathological behaviors.

2) We prove sample complexity bounds for deep transformer-based LLMs pre-trained on *non-iid* data. They are more general than existing results (see Table 1 for a full comparison with the literature) and are verified in practice on the Llama and Gemma herd of models from 2B to 27B.

3) We derive ICL generalization bounds when LLMs are prompted with Markov chains. We validate our results by showing that the most recent open-source LLMs obey the in-context *scaling laws* predicted by our theory.

Organization of the paper. We first briefly introduce the background material on transformer-based autoregressive LLMs and Markov chains before presenting our main contribution that formalizes the multi-step inference mechanism of a pre-trained LLM as a Markov chain. We leverage this formulation to derive sample complexity and generalization bounds on large language models in pre-training and in-context learning which we validate in practice on commonly used LLMs. Due to space constraints, we mostly present the Markov chain formulation in this summary and refer the reader to the full version of the paper in Zekri et al. (2025) for more details.

2 Background Knowledge

We recall some basic notions on Markov chains and LLMs.

Markov chains. Let Ω be a discrete finite set of size $|\Omega|$. A discrete-time, time-homogeneous Markov chain $\mathrm{MC}(\Omega, \mathbf{Q})$ defined on a state space $\Omega = \{x_i\}_{i=1}^{|\Omega|}$ with transition matrix $\mathbf{Q} \in \mathbb{R}^{|\Omega|\times|\Omega|}$ with entries $\mathbf{Q}_{ij} = \mathbf{Q}(x_i, x_j) \in [0, 1]$ is a sequence of random variables $(\mathbf{X}_1, \mathbf{X}_2, \ldots)$ taking values in Ω such that for any $n \in \mathbb{N}$ and $(x_1, \ldots, x_{n+1}) \in \Omega^{n+1}$, we have

$$\begin{aligned}\mathbb{P}(\mathbf{X}_{n+1} = x_{n+1} \mid \mathbf{X}_n = x_n, \ldots, \mathbf{X}_1 = x_1) \\ = \mathbb{P}(\mathbf{X}_{n+1} = x_{n+1} \mid \mathbf{X}_n = x_n) \\ \coloneqq \mathbf{Q}(x_n, x_{n+1}).\end{aligned}$$

A distribution π on Ω is called stationary if $\pi\mathbf{Q} = \pi$. For any $x \in \Omega$, $\mathrm{MC}(\Omega, \mathbf{Q})$ converges to π if $\lim_{n\to\infty} d_{\mathrm{TV}}(\mathbf{Q}^n(x, \cdot), \pi) = 0$, where $\mathbf{Q}^n(x, \cdot)$ denotes the probability of $\mathbf{X}_n$ conditioned on $\mathbf{X}_1 = x$ and the total variation between two distributions $\mathbb{P}$ and $\mathbb{Q}$, defined on $(\Omega, \mathcal{F})$, is $d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \coloneqq \sup_{A\in\mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|$.

Large language models. Let $\mathcal{V}$ denote a dictionary of size T, referred to as the *vocabulary size*, used to encode an arbitrary sequence into a sequence of predefined tokens belonging to $\mathcal{V}$. We assume that our model admits a maximum of K tokens as input, referred to as the *context window* of the model. The domain of the transformer-based autoregressive LLM is the set of all sequences consisting of at most K elements of $\mathcal{V}$. We denote this space by

$\mathcal{V}_K^*$, i.e., $\mathcal{V}_K^* := \{v \in \mathcal{V}^*, |v| \leq K\}$ with $|v|$ the length of v. We define an LLM with trainable parameters $\boldsymbol{\Theta}$ as a function $f_{\boldsymbol{\Theta}}^{T,K} : \mathcal{V}_K^* \to \Delta(\mathcal{V})$, where $\Delta(\mathcal{V})$ is the probability simplex over $\mathcal{V}$. Given a sequence of tokens v, $f_{\boldsymbol{\Theta}}^{T,K}$ outputs a probability distribution over the whole state space indicating the likelihood for each of its elements to appear after v. Noting that an LLM is defined for a given vocabulary size T and context window K, we will drop the exponents and simply write $f_{\boldsymbol{\Theta}}$ to ease the notations. Without loss of generality, $f_{\boldsymbol{\Theta}}$ is assumed to be a transformer model with L layers and H heads, consisting of alternating multi-head attention (MHA) and feed-forward blocks. We consider a setting where the learner's objective is to approximate the probability distribution of sequences over an input vocabulary $\mathbb{P}_{\mathcal{L}} : \mathcal{P}(\mathcal{V}_K^*) \to [0, 1]$ where $\mathcal{P}(\mathcal{V}_K^*)$ denotes the powerset of $\mathcal{V}_K^*$.

3 Large Language Models as Markov Chains

To proceed, we formally define a Markov chain that explicitly captures the inference of a given LLM $f_{\boldsymbol{\Theta}}$. We build upon a high-level idea that associates a tokenized input sequence with a state v_i, from which we transition to a new state $v_j = [v_i, v]$ by concatenating the token v predicted by an LLM to it. We then provide a theoretical characterization of this Markov chain. Our analysis holds for any model with finite fixed vocabulary, and context window. This setup includes deep transformer-based LLMs while excluding models such as RNNs and LSTMs.

Markov chain formalization. We first introduce the notion of "incompatible sequences" in the sense of next-token prediction.

Definition 3.1 (Incompatible Sequences)**.** *Let $u, v \in \mathcal{V}_K^*$ be two sequences of at most K tokens. We say that v is incompatible with u if v cannot be a plausible completion of u, that is*

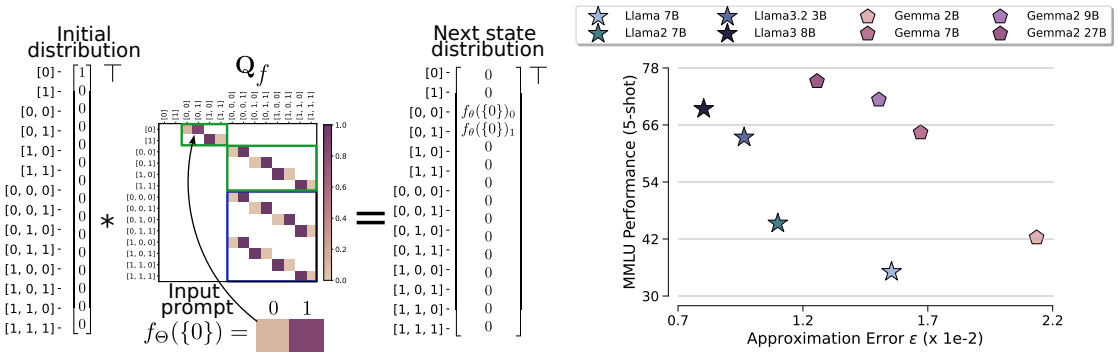
$$\exists l \in \{1, \dots, |u| - 1\}, s.t.\ (u)_{l+1} \neq (v)_l. \tag{1}$$

$\mathcal{I}$ denotes the set of incompatible sequences, i.e., $\mathcal{I} \coloneqq \{(u, v)\ s.t.\ v \text{ is incompatible with } u\}$.

The order matters since a necessary condition for v to be compatible with u is that v has as many or one more token than u. We now proceed with the definition of the transition matrix associated with an LLM $f_{\boldsymbol{\Theta}}$.

Proposition 3.2. *Any large language model $f_{\boldsymbol{\Theta}}$ can be equivalently represented by a Markov chain $MC(\mathcal{V}_K^*, \mathbf{Q}_f)$, with a sparse transition matrix $\mathbf{Q}_f \in \mathbb{R}^{|\mathcal{V}_K^*| \times |\mathcal{V}_K^*|}$ that writes for any $v_i, v_j \in \mathcal{V}_K^*$:*

$$\mathbf{Q}_f(v_i, v_j) = \begin{cases} 0, & \text{if } (v_i, v_j) \in \mathcal{I} \\ \{f_{\boldsymbol{\Theta}}(v_i)\}_{j_0}, & \text{otherwise,} \end{cases}$$



(a) Illustration of Proposition 3.2.

(b) LLMs' Sample Complexity.

Figure 2: (a) *Proposition 3.2 with* $T = 2$ *and* $K = 3$. (b) We plot the Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) performance with respect to the approximation error ϵ predicted by our theory. Using the real number of pre-training tokens used in the Llama and Gemma families (Dubey et al., 2024; Riviere et al., 2024), we show that our predicted approximation error ϵ correlates with the real performance, with different trends between the models' families.

where j_0 *denotes the index in* $\mathcal{V}$ *of the last token of* v_j*. We have* $|\mathcal{V}_K^*| = T(T^K - 1)/(T - 1)$ *and the proportion of non-zero elements in* $\mathbf{Q}_f$ *is* $(T - 1)/(T^K - 1)$*.*

Proposition 3.2 states that if v_j is incompatible with v_i, then the probability to go from the state v_i to v_j is zero as LLM only transitions between compatible sequences. For such compatible sequences, the probability of going from v_i to v_j is naturally the next-token probability associated with the last token of v_j, i.e., $\{f_{\mathbf{\Theta}}(v_i)\}_{j_0}$.

Example: binary sequences. To help unpack Proposition 3.2, we illustrate it with $T = 2$ and $K = 3$ in Fig. 2(a) (see the general case in Fig. 1). Here, we start with an input prompt consisting of one token. We then transition to sequences of increased size, while attributing a 0 probability to incompatible sequences. Each generation step for different values of $k < K$ defines green rectangular blocks of size $T^k \times T^{k+1}$ in the transition matrix $\mathbf{Q}_f$. Reaching the blue *square* block in the transition matrix means that the input sequence is of the maximum size K. The model can no longer append tokens to it and has to delete the first token to proceed. This blue block is of size $T^K \times T^K$: it captures transitions between sequences of the maximum admissible length. We define similarly the reference transition matrix $\mathbf{Q}^*$ of the language where the probability of transitions $\{f_{\mathbf{\Theta}}(v_i)\}_j$ are replaced by ground-truth probabilities $\mathbb{P}_{\mathcal{L}}(v_j \mid v_i)$. To use $\mathbf{Q}_f$ as $f_{\mathbf{\Theta}}$, it is sufficient to define an input distribution δ_0 of the Markov chain based on input prompt v. It is a one-hot encoding vector of size $|\mathcal{V}_K^*|$ with 1 at the position of the state corresponding to v. Then, the transition to the next state writes $\delta_1 = \delta_0 \mathbf{Q}_f$. The output of $f_{\mathbf{\Theta}}(v)$ for individual tokens in $\mathcal{V}$ would then correspond to probabilities in δ_1 for states that are concatenations of v with T tokens from $\mathcal{V}$. This process is illustrated in Fig. 2(a). In the full version of the paper (see Zekri et al. (2025)), we study the properties of $\mathrm{MC}(\mathcal{V}_K^*, \mathbf{Q}_f)$ and discuss further implications of such formulation regarding LLMs' behavior in practice.

Method	Pre-Train.	ICL	Input	Model-Dep.	Exp. val.
Xie et al. (2022)	×	✓	HMM	×	✓
Zhang et al. (2023b)	✓	✓	MC	✓	×
Li et al. (2023)	✓	✓	MC	✓	✓
Lotfi et al. (2024)	✓	×	non-iid	×	✓
Ours	✓	✓	non-iid	✓	✓

Table 1: LLMs generalization bounds proposed in the literature. **Pre-train.** stands for pre-training; **Input** refers to the assumptions on the data generating process (Markov Chains (MC); Hidden Markov Model (HMM); **Model-dep.** means explicit dependence on model's architecture; **Exp.val.** stands for the experimental validation of the theory.

Sample complexity and generalization. The Markov chain equivalence enables us to study the generalization of an LLM $f_{\boldsymbol{\Theta}}$ as its capacity to infer correctly all the elements of $\mathbf{Q}_f$, while approximating the true reference matrix of transition probabilities $\mathbf{Q}^*$. The hardness of this task is highlighted by the fact that an LLM observes a negligible amount of $\mathbf{Q}^*$'s elements during its pre-training. Indeed, for GPT-3 (Brown et al., 2020), this represents 5×10^{11} training tokens, which pales in comparison with the number of non-zero elements in $\mathbf{Q}_f$, given by $T^{K+1} \approx 10^{9632}$. In our work, we derived the sample complexity of LLMs, that is the minimal number of pre-training tokens needed to be ϵ close to the reference matrix. The proof relies on our novel pre-training generalization bound on LLMs that use concentration inequalities of dependent random variables proposed in Paulin (2015). We extend this bound to the in-context learning framework and show that our theoretical results correctly capture the behavior of LLMs in practice, e.g., in Fig. 2(b). We summarize in Table 1 the novelty of our generalization bounds.

4 Conclusion

This paper proposed an explicit characterization of the inference mechanism in large language models through an equivalent finite-state Markov chain. We provided an insightful theoretical analysis based on the established characterization and the ability of the LLM to infer the transition kernel approximating the true transition probabilities of language. Experiments on commonly used LLMs validate our findings. We adapted our results to in-context learning where experiments confirm our theoretical insights. In the future, we hope that the proposed equivalence will have far-reaching implications on our understanding of LLMs and allow for a more fine-grained understanding of their expressiveness.

References

Brown, T., Mann, B., Ryder, N., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Edelman, B. L., Edelman, E., Goel, S., Malach, E., and Tsilivis, N. The evolution of statistical induction heads: In-context learning markov chains. *arXiv preprint arXiv:2402.11004*, 2024.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Ildiz, M. E., Huang, Y., Li, Y., Rawat, A. S., and Oymak, S. From Self-Attention to Markov Models: Unveiling the Dynamics of Generative Transformers. *arXiv preprint arXiv:2402.13512*, 2024.

Jeon, H. J., Lee, J. D., Lei, Q., and Van Roy, B. An Information-Theoretic Analysis of In-Context Learning. In *International Conference on Machine Learning*, volume 235, pp. 21522–21554, 2024.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pp. 19565–19594. PMLR, 2023.

Lotfi, S., Finzi, M., Kuang, Y., Rudner, T. G., Goldblum, M., and Wilson, A. G. Non-vacuous generalization bounds for large language models. *arXiv preprint arXiv:2312.17173*, 2023.

Lotfi, S., Kuang, Y., Amos, B., Goldblum, M., Finzi, M., and Wilson, A. G. Unlocking tokens as data points for generalization bounds on larger language models. *arXiv preprint arXiv:2407.18158*, 2024.

Makkuva, A. V., Bondaschi, M., Girish, A., Nagle, A., Jaggi, M., Kim, H., and Gastpar, M. Attention with Markov: A framework for principled analysis of transformers via markov chains. *arXiv preprint arXiv:2402.04161*, 2024.

Paulin, D. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Probab.*, 20:79, 2015.

Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., and et al., J. F. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Roberts, A., Raffel, C., and Shazeer, N. How much knowledge can you pack into the parameters of a language model? In *Empirical Methods in Natural Language Processing*, pp. 5418–5426. Association for Computational Linguistics, 2020.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An Explanation of In-context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations*, 2022.

Zekri, O., Odonnat, A., Benechehab, A., Bleistein, L., Boullé, N., and Redko, I. Large language models as markov chains, 2025. URL https://arxiv.org/abs/2410.02724.

Zhang, W., Deng, Y., Liu, B., Pan, S. J., and Bing, L. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*, 2023a.

Zhang, Y., Zhang, F., Yang, Z., and Wang, Z. What and how does in-context learning learn? Bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023b.

Data Complexity is not Data uncertainty

Timothee Mickus[1], Aman Sinha*[2], and Raúl Vazquéz[1]

[1]University of Helsinki – Finlande
[2]Université de Lorraine – Doctorant – France

Résumé

La difficulté intrinsèque d'un exemple, enracinée dans son ambiguïté inhérente, est un facteur clé souvent négligé dans l'évaluation des modèles neuronaux en traitement automatique du langage (TAL). Nous étudions l'interaction et la divergence entre diverses métriques d'évaluation de cette difficulté intrinsèque, notamment le désaccord entre annotateurs, la dynamique d'entraînement et la confiance des modèles. À travers une analyse approfondie portant sur 29 modèles et trois jeux de données, nous montrons que, bien que ces métriques soient corrélées, leurs relations ne sont ni linéaires ni monotones. En démêlant ces différentes dimensions de l'incertitude, nous visons à affiner notre compréhension de la complexité des données et de ses implications pour l'évaluation et l'amélioration des modèles de TAL.

Mots-Clés: TAL, complexité des données, variation inter, annotateurs

*Intervenant

Transport optimal

Procrustes-Wasserstein for isometric invariant interpolations

Davide Adamo [1,2] & Marco Corneli [1,2]

[1] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, Fr*

[2] *Université Côte d'Azur, UMR 7264 CEPAM, CNRS, Nice, France*

davide.adamo@inria.fr

Résumé. Les problèmes de transport optimal (OT) représentent des techniques puissantes pour découvrir des correspondances entre des ensembles d'objets de manière totalement non supervisée. En particulier, lorsque les objets sont représentés sous forme de nuages de points, la distance de Wasserstein s'est avérée très efficace pour quantifier leur similarité tout en tenant compte des propriétés géométriques et de distribution des données d'entrée. Toutefois, les limites de la distance de Wasserstein sont bien connues. Sa sensibilité à l'intégration des nuages de points dans l'espace, et donc aux isométries, la rend inadaptée à certaines applications, telles que la mise en correspondance des nuages de points. Pour contourner ce problème, la distance de Procrustes-Wasserstein (PW) est une extension notable invariante par rapport aux isométries. L'invariance par rapport aux transformations rigides telles que les rotations et les réflexions fait de la PW une métrique OT plus appropriée pour résoudre des tâches telles que l'alignement et la comparaison de nuages de points. Dans ce travail, nous présentons les barycentres de PW comme une méthode robuste pour le calcul d'interpolations représentatives qui tiennent compte des caractéristiques géométriques globales et locales des données. En outre, nous étendons cette méthode à une version détendue du problème, ce qui permet de calculer des barycentres robustes.

Mots-clés. Transport optimal, Analyse procustéenne, Nuage de points.

Abstract. Optimal transport (OT) problems provide powerful techniques for discovering correspondences between sets of objects in a fully unsupervised way. In particular, when objects are represented as point clouds, the Wasserstein distance has proven to be highly effective in quantifying their similarity while accounting for both geometric and distributional properties of the input data. However, the limitations of the Wasserstein distance are well known. Its sensitivity to the embedding of point clouds in space, and thus to isometries, makes it unsuitable for certain applications, such as point cloud matching. To circumvent this problem, a notable isometry-invariant extension is represented by the Procrustes-Wasserstein distance (PW). The invariance with respect to rigid transformations such as rotations and reflections, makes the PW a more suitable OT metric for solving tasks like point cloud alignment and comparison. In this work we introduce the PW barycenters as a robust method for the computation of representative interpolations that account for both global and local geometric characteristics of the data. Moreover, we extend this method to a relaxed version of the problem, enabling the computation of robust barycenters.

Keywords. Optimal transport, Procrustes Analysis, Point clouds.

1 PW background and barycenter(s) definition

In order to generalize classical optimal transport problems to account for global transformations, such as rotations and reflections in the space, Zhang et al. (2017) and Grave et al. (2019) introduced a mix of Wasserstein and Procrustes analysis, known as Procrustes-Wasserstein cost. PW aims at jointly estimating an orthogonal and a permutation matrix to solve tasks like objects alignment. Despite the heterogeneous use of PW problems in the different domains, in Adamo et al. (2025) we carefully define the quotient space of discrete probability measures over which the PW defines a distance. Moreover, we extend the OT literature introducing the discrete PW barycenters. In the following sections we resume these definitions.

Denote by Σ_n the $n-1$ probability simplex. So when saying that $\mathbf{p} := (p_1, \dots, p_n) \in \Sigma_n$, we mean $p_i \geq 0$ for all i and $\sum_{i=1}^n p_i = 1$. Denote also by $\langle \cdot, \cdot \rangle_F$ the Frobenius dot product and $\mathcal{O}(d)$ the set of the orthogonal matrices of order d.

PW distance. Consider two matrices $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^{m \times d}$, where $\mathbf{x}_i$ (respectively $\mathbf{x}^j$) is the i-th row (j-th column) of X. Similarly for Y. We attach two discrete probability measures μ_X and μ_Y to X and Y, respectively:

$$\mu_X = \sum_{i=1}^{n} p_i \delta_{\mathbf{x}_i} \quad \text{and} \quad \mu_Y = \sum_{j=1}^{m} q_j \delta_{\mathbf{y}_j}, \qquad \mathbf{p} \in \Sigma_n \quad \text{and} \quad \mathbf{q} \in \Sigma_m.$$

The Procrustes-Wasserstein distance of order 2 is defined by

$$PW_2(\mu_X, \mu_Y) := \left(\min_{\substack{P \in \mathcal{O}(d) \\ \Gamma \in \Pi(\mathbf{p}, \mathbf{q})}} \langle C_P(X, Y), \Gamma \rangle_F \right)^{1/2}, \tag{1}$$

where $(C_P(X,Y))_{ij} = \|\mathbf{x}_i - \mathbf{y}_j P\|_2^2$ and $\Pi(\mathbf{p}, \mathbf{q})$ is the set of the admissible transport plans.

$PW_2(\cdot, \cdot)$ defines a distance on the space of discrete probability measures that are equivalent up to a permutation of their probability vectors and a rigid transformation (rotation, reflection, or a combination of the two) of their supports (Adamo et al., 2025, Theorem 2.1).

PW barycenter(s). Consider now r empirical measures $\mu_{X_1}, \dots, \mu_{X_r}$, with supports $\{X_j\}_{j=1}^r$ and probability vectors $\{\mathbf{p}_j\}_{j=1}^r$. We seek for an unknown measure μ_X with support $X \in \mathbb{R}^{n \times d}$ and weights $\mathbf{p}$ given by the minimizer of the following problem

$$f(\mathbf{p}, X) := \sum_{j=1}^{r} \lambda_j PW_2^2(\mu_X, \mu_{X_j}), \tag{2}$$

where λ_j are positive weights associated to each measure μ_{X_j}, with $\boldsymbol{\lambda} := (\lambda_1, \dots, \lambda_r) \in \Sigma_r$.

Since Problem (2) is not convex w.r.t. the barycenter locations, by assuming the probability vector $\mathbf{p}$ known, which is often user-defined, we can resort to a minimization of a local quadratic approximation of each $PW_2^2(\mu_X, \mu_{X_j})$ at X. Thus, the Newton update reads

$$X \leftarrow \text{diag}(\mathbf{p}^{-1}) \sum_{j=1}^{r} \lambda_j \Gamma_j^* X_j P_j^*. \tag{3}$$

This update formula can be interpreted as follows. For each j the matrix Γ_j^* computes weighted barycenters of points in X_j, with weights defined by the optimal transport plan. Analogously to the Wasserstein barycenter update (Cuturi and Doucet, 2014), where each point in X_j contributes to the updated locations in X proportionally to Γ_j^*. However, in the PW framework, the additional right multiplication by P_j^* allows for a simultaneous optimal alignment of the barycenter to X_j. The numerical schemes for the optimisation of the PW problem and for the PW barycenter computation are reported in Algorithm 1 and Algorithm 2 of (Adamo et al., 2025).

2 Regularized PW barycenter(s)

As is common in the OT community, in this section, we extend the PW barycenter problem by considering an entropic regularization. Consider the following relaxed version of the PW problem

$$PW_{\epsilon,2}^2(\mu_X, \mu_{YP}) = \min_{\substack{P \in \mathcal{O}(d) \\ \Gamma \in \Pi(\mathbf{p},\mathbf{q})}} \langle C_P(X,Y), \Gamma \rangle_F + \epsilon H(\Gamma), \tag{4}$$

where ϵ is a non-negative parameter controlling the strength of the regularization and $H(\Gamma) := \sum_{i,j} \Gamma_{ij} \log(\Gamma_{ij})$ is the entropy.

Searching now for a measure μ_X with unknown support $X \in \mathbb{R}^{n \times d}$ and weights $\mathbf{p}$, given r measures $\mu_{X_1}, \ldots, \mu_{X_r}$, translates into the following minimization problem

$$f_\epsilon(\mathbf{p}, X) := \sum_{j=1}^{r} \lambda_j PW_{\epsilon,2}^2(\mu_X, \mu_{X_j}). \tag{5}$$

The resolution of this problem follows a similar approach to the classical case. Given the probability vector $\mathbf{p}$ and assuming (Γ_j^*, P_j^*) are the solutions of the regularized PW problem $PW_{\epsilon,2}^2(\mu_X, \mu_{X_j})$, the Newton update is still given by Eq. (3). Notably, at the optimum, both the gradient and the Hessian of Eq. (5) are independent of the regularization term $H(\cdot)$. The optimization scheme for the computation of the barycenter is equivalent to Algorithm 2 in (Adamo et al., 2025), where in this case we solve each PW sub-problem independently using the Sinkhorn algorithm (Cuturi, 2013). Under this framework, the optimization of Eq. (5) can be seen as a particular case of the one discussed in Alvarez-Melis et al. (2019).

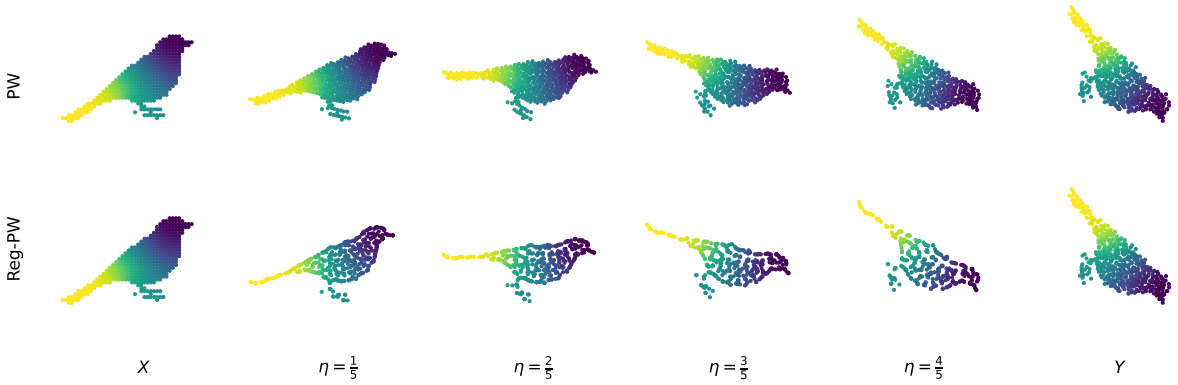


Figure 1: Progressive interpolations of two point clouds describing a bird in different positions. The four central columns of the grid correspond to PW barycenters, each reflecting a specific interpolation step (defined by η). The two rows represent barycenters calculated using the classical formulation (PW) and considering an entropic relaxation of the problem (Reg-PW).

3 Point clouds application

Point cloud matching represents a fundamental task in computational geometry, computer vision and shape analysis. In this context, the PW distance and the PW barycenters offer robust solutions for the computation of representative objects that account for rigid transformations in the space. To illustrate the effectiveness of the proposed barycenter, we present a 2D toy example where we consider two measures μ_X and μ_Y, with associated locations X and Y, representing different instances of the "same" object. Specifically, X represents a point cloud depicting a bird in a particular pose. We define Y as a modified version of X with the following transformations: addition of Gaussian perturbations and random vertex permutation; addition of extra vertices; application of a random rotation. Within this setting, our goal is to generate interpolated shapes that progressively move from X to Y. Set thus $\eta \in [0, 1]$ and re-write $\lambda_X = 1 - \eta$ and $\lambda_Y = \eta$. By varying η we can iterate the minimization problem and model the intermediate stages. We compute PW barycenters using both the classical problem formulation (Problem 2) and the regularized version (Problem 5).

In Figure 1 we can follow the progressive interpolations between the two point clouds. The four central columns correspond to PW barycenters at different interpolation steps, while the two rows compare the classical PW barycenter (top) with the relaxed version (bottom). The colors of the barycenters are given by transporting the colors of X via the optimal plan Γ_X^*. This highlights how the cloud evolves while maintaining the structural consistency of the original shape.

(Extra 2D/3D examples and barycenter properties will be discussed at the oral)

References

Adamo, D., Corneli, M., Vuillien, M., and Vila, E. (2025). An in depth look at the Procrustes-Wasserstein distance: properties and barycenters. working paper or preprint.

Alvarez-Melis, D., Jegelka, S., and Jaakkola, T. S. (2019). Towards optimal transport with global invariances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1870–1879. PMLR.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Cuturi, M. and Doucet, A. (2014). Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR.

Grave, E., Joulin, A., and Berthet, Q. (2019). Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR.

Zhang, M., Liu, Y., Luan, H., and Sun, M. (2017). Earth mover's distance minimization for unsupervised bilingual lexicon induction. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Optimal Transport-Based Domain Adaptation for Rotated Linear Regression

Brian Britos [1] & Mathias Bourel [2]

[1] *Aix-Marseille Université, France, brian.BRITOS-SIMMARI@etu.univ-amu.fr*
[2] *Universidad de la República, Uruguay, mbourel@fing.edu.uy*

Résumé. Le transport optimal (OT) s'est révélé efficace pour l'adaptation de domaine (DA) en alignant les distributions entre domaines aux propriétés statistiques différentes. S'appuyant sur l'approche de Courty et al. (2016), qui proposaient de projeter les données sources vers le domaine cible afin d'améliorer le transfert de modèles, nous étudions ici un problème supervisé d'adaptation de domaine impliquant des modèles de régression linéaire soumis à des rotations. Ce travail en cours concerne des situations où les domaines source et cible sont liés par une rotation, un cas fréquent dans des contextes tels que l'étalonnage de capteurs ou les changements d'orientation d'images.

Nous montrons que dans $\mathbb{R}^2$, lorsqu'on utilise un coût de type p-norme avec $p \geq 2$, le transport optimal permet de retrouver la rotation sous-jacente. À partir de ce résultat théorique, nous proposons un algorithme combinant un regroupement par K-moyennes, OT, et une décomposition en valeurs singulières (SVD) pour estimer l'angle de rotation et adapter le modèle de régression. Cette méthode est particulièrement efficace lorsque les données du domaine cible sont limitées, car elle exploite la richesse des données du domaine source pour améliorer la généralisation. Nos contributions offrent ainsi un éclairage à la fois théorique et pratique sur l'utilisation du transport optimal pour l'adaptation de modèles sous des transformations géométriques.

Mots-clés. Transport Optimal, Adaptation de Domaine, Régression Linéaire

Abstract. Optimal Transport (OT) has proven effective for domain adaptation (DA) by aligning distributions across domains with differing statistical properties. Building on the approach of Courty et al. (2016), who mapped source data to the target domain for improved model transfer, we focus on a supervised DA problem involving linear regression models under rotational shifts. This ongoing work considers cases where source and target domains are related by a rotation—common in applications like sensor calibration or image orientation.

We show that in $\mathbb{R}^2$, when using a p-norm cost with $p \geq 2$, the optimal transport map recovers the underlying rotation. Based on this, we propose an algorithm that combines K-means clustering, OT, and singular value decomposition (SVD) to estimate the rotation angle and adapt the regression model. This method is particularly effective when the target domain is sparsely sampled, leveraging abundant source data for improved generalization. Our contributions offer both theoretical and practical insights into OT-based model adaptation under geometric transformations.

Keywords. Optimal Transport, Domain Adaptation, Linear Regression

1 Introduction

Optimal Transport (OT), introduced by Monge (1781) and later formalized by Kantorovich (1942), provides a mathematical framework for efficiently transforming one probability distribution into another. Peyré et al. (2019) offer a detailed overview of its computational aspects, making it a practical tool for various applications. In domain adaptation (DA), OT has proven particularly effective in aligning source and target distributions, especially when labeled data in the target domain are scarce.

In this work, we investigate a supervised DA problem where the source and target domains in $\mathbb{R}^2$ are related by an unknown rotation. We leverage OT to estimate the rotation angle, facilitating the adaptation of a simple linear regression model from the source to the target domain. Our approach combines K-means clustering, OT-based alignment, and a singular value decomposition (SVD) method for robust angle estimation. This framework is particularly useful when target data is limited, allowing the source domain to guide adaptation effectively.

2 Our method

We begin this section by summarizing two crucial technical tools for our method.

- Let $a, b \in \mathbb{R}$ be the coefficients of a line r in the plane $\mathbb{R}^2$. Then, the coefficients of the line obtained by rotating r by an angle θ around the origin are:

$$\tilde{a} = \frac{a\cos\theta + \sin\theta}{\cos\theta - a\sin\theta}, \quad \tilde{b} = b(\cos\theta + a\sin\theta). \tag{1}$$

- The algorithm proposed in Arun et al. (1987) to estimate the rotation angle between two sets in the plane $X^s = \{(x_1^s, y_1^s), \ldots, (x_n^s, y_n^s)\}$ and $X^t = \{(x_1^t, y_1^t), \ldots, (x_n^t, y_n^t)\}$.

2.1 Principal theoretical result

Below we present a novel result concerning a geometric property of optimal transport in the plane. This result will later play a critical role in the design for an algorithm for domain adaptation within the same framework.

Proposition: Let $X^s = \{(x_1^s, y_1^s), \ldots, (x_n^s, y_n^s)\} \subset \mathbb{R}^2$ and $X^t = \{(x_1^t, y_1^t), \ldots, (x_n^t, y_n^t)\} \subset \mathbb{R}^2$ such that $|X^s| = |X^t| = n$, and related by a rotation R_θ of angle θ, that is, each (x_i^t, y_i^t) is of the form $R_\theta(x_i^s, y_i^s)$. Let μ and ν denote the empirical uniform measures over X^s and X^t, respectively, and assume the cost function is the p-norm:

$$c((x_i^s, y_i^s), (x_j^t, y_j^t)) = \|(x_j^t, y_j^t) - (x_i^s, y_i^s)\|_p, \quad p \geq 2.$$

Then, the optimal transport map T is unique and aligns with the rotation R_θ in the sense that $T((x_i^s, y_i^s)) = R_\theta(x_i^s, y_i^s)$ for all $i = 1, \ldots, n$.

Now we shift our attention to the following domain adaptation problem: we have two domains, $\mathcal{D}^s = \{(x_1^s, y_1^s), \ldots, (x_{n_s}^s, y_{n_s}^s)\}$ and $\mathcal{D}^t = \{(x_1^t, y_1^t), \ldots, (x_{n_t}^t, y_{n_t}^t)\}$, in $\mathbb{R}^2$, where $n_s \gg n_t$. We assume that the underlying structure of the two domains are lines related by a unknown rotation R_θ, as shown in Step 0 of Figure 1. Here, $\mathcal{D}^s$ serves as a training set and $\mathcal{D}^t$ as a development set, modelling scenarios where training data is abundant but target data is scarce due to rarity, cost, or physical constraints. This setup reflects challenges associated with data drift, common in machine learning (see Sahiner et al. 2023). The goal is to adapt a linear regression model, $y = a_s x + b_s$, trained on $\mathcal{D}^s$, to perform effectively on $\mathcal{D}^t$. The procedure is detailed in the algorithm 2.

Using the above results together with $K-$means we can estimate the angle rotation θ as follows: first using K-means, with $K = n_t$, we construct a set of centroids $\mathcal{D}^k = \{(x_1^k, y_1^k), \ldots, (x_{n_t}^k, y_{n_t}^k)\}$ (shown as purple dots in Step 1 of Figure 1). Next, taking uniform measures over these sets and using the p-norm as cost function allow us to approximate the assumptions of the previous proposition. Although $\mathcal{D}^t$ and $\mathcal{D}^k$ are not strictly related by a rotation, the optimal transport map T (right of Figure 1) approximates the rotation. Finally we apply the proposed algorithm in Arun et al. (1987) to estimate the rotation angle θ. This procedure is summarized in algorithm 1.

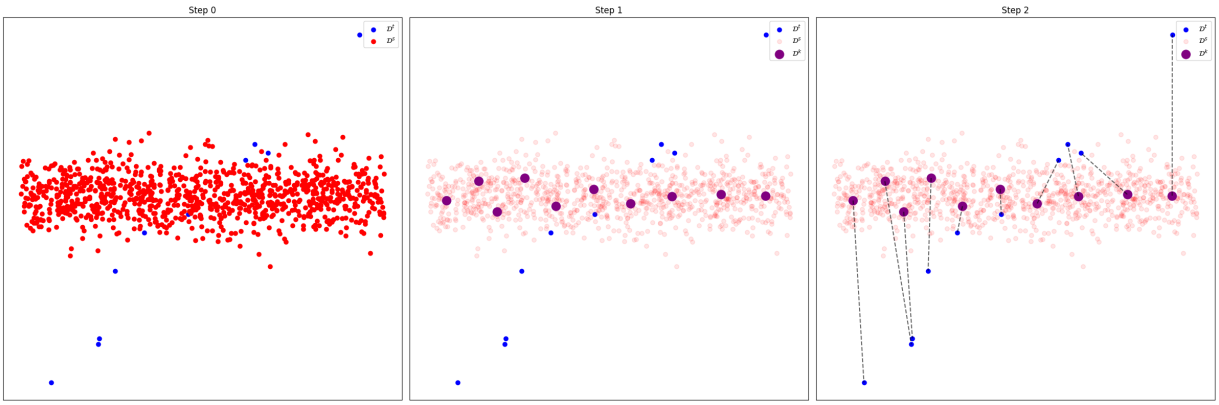


Figure 1: Red points are the source set, blue ones the target set and the purple dots are the K-means centroids. Step 0 is the raw data, in Step 1 we apply K-means with $K = n_t$ and in Step 3 we find the optimal transport plan.

Algorithm 1 Angle Estimation Using K-Means, Optimal Transport, and SVD

Require: $\mathcal{D}^s, \mathcal{D}^t$: Source and target domain.
Ensure: $\hat{\theta}$: Estimated rotation angle.
Step 1: $\mathcal{D}^k \leftarrow \text{KMeans}(\mathcal{D}_s, \text{n_clusters} = n_t)$.
Step 2: $T \leftarrow$ Optimal transport plan obtained with uniform measures μ, ν and $||\cdot||_p$.
Step 3: Use the optimal transport plan T to sort the sets $\mathcal{D}^t$ and $\mathcal{D}^k$.
Step 4: $\hat{\theta} \leftarrow$ Estimate rotation angle between $\mathcal{D}^s$ and $\mathcal{D}^k$ using Arun K. S. et al. (1987).
return $\hat{\theta}$

Algorithm 2 Adapted Linear Regression in $\mathbb{R}^2$

Require: $\mathcal{D}^s, \mathcal{D}^t$: source and target domains; N: number of repetitions; p: bootstrap proportion.
Ensure: a_r, b_r: adapted regression coefficients.
  Fit linear regression on $\mathcal{D}^s$ to obtain a_s, b_s
  for $i = 1$ to N **do**
    $\mathcal{D}^k \leftarrow$ Sample bootstrap subset $\mathcal{D}^b$ (proportion p).
    $\hat{\theta} \leftarrow$ Apply Algorithm 1 to $\mathcal{D}^t$ and $\mathcal{D}^b$.
    $a_i, b_i \leftarrow$ Apply equation 1 with a_s, b_s, and $\hat{\theta}$.
  end for
  return $a_r = \text{median}(\{a_1, \ldots, a_N\}), b_r = \text{median}(\{b_1, \ldots, b_N\})$

The motivation for including a bootstrapping step prior to estimating the rotation angle is to ensure the generation of distinct centroids during the application of K-means. Additionally, at the end we employ the median instead of the mean, as the procedure has significant outliers.

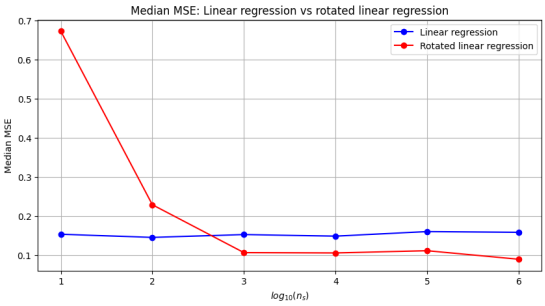
3 Simulations

In this section we present two simulations to evaluate when the proposed procedure outperforms fitting the regression model solely on the target domain. These tests illustrate its practical advantages and highlight scenarios of significant predictive improvement. We explore two distinct scenarios: in 3.1, we vary the source domain size n_s while fixing the angle, noise level, and target domain size n_t whereas in 3.2, we vary the angle and noise level, keeping n_s and n_t fixed.

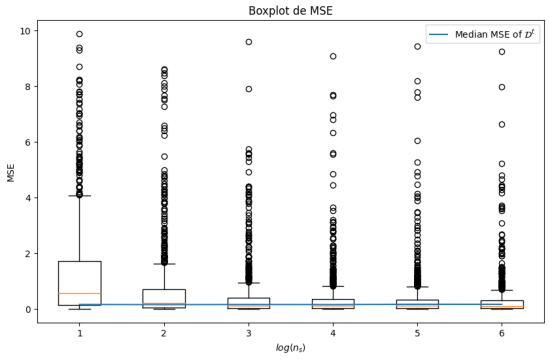
The simulations were carried out by drawing points from a straight line and adding Gaussian noise, that means, fixing $a \in \mathbb{R}$ and the noise variance $\sigma > 0$ we generate the data set following $y = ax + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma)$. As we always can apply a rotation to the plane so the source domain is in the x-axis, we generate the source domain as $y = \epsilon$ (i.e. $a = 0$). On the other hand, fixing an angle θ, we generate the target domain following $y = \arctan(\theta)x + \epsilon$ where the noise variance σ is the same for both.

3.1 Varying the cardinality of $\mathcal{D}^s$

We simulate data with a fixed rotation angle $\theta = \frac{\pi}{4}$, target domain size $n_t = 10$, noise level $\sigma = 1$ and vary the source domain size n_s. For each n_s, the procedure is applied 1000 times, and the median Mean Square Error (MSE) is recorded (Figure 2a). When n_s and n_t are similar (leftmost points), the procedure shows no benefit. However, as the ratio $\frac{n_s}{n_t}$ increases, our method outperforms and exhibits asymptotic behaviour. While the proposed method exhibits a high number of outliers, the median suggests it provides more reliable results in most cases.



(a) The target domain size n_t is fixed at 10, while the source domain size n_s varies in $\{10, 10^2, 10^3, 10^4, 10^5, 10^6\}$. Blue dots show the median MSE for regression on $\mathcal{D}^t$, and red points show the median MSE using our method.



(b) Boxplots of the same experiment as Figure 2a. The blue line represents regression on $\mathcal{D}^t$. Our method shows more outliers but better median performance.

Figure 2: Comparison of the performance of our method against the standard linear regression across different settings of n_s.

3.2 Varying the angle θ and noise variance σ

To assess where our method performs best, we fix $n_s = 1000$ and $n_t = 50$, and evaluate variation over the θ–σ plane, varying $\theta \in \frac{5}{6}\pi, \ldots, \frac{1}{6}\pi$ and $\sigma \in 0.1, 0.2, 0.5, 1, 2, 5$. We defined the variation as $\frac{MSE}{MSE_R} - 1$, where MSE is the mean square error of the linear regression obtained using only the target domain $\mathcal{D}^t$ and MSE_R is the mean square error of the regression obtained by our method Figure 3 shows the median result over 100 runs. Performance improves with higher noise, leveraging the larger source data. Results are also better for θ near $\frac{\pi}{2}$, likely because linear regression struggles with near-vertical slopes.

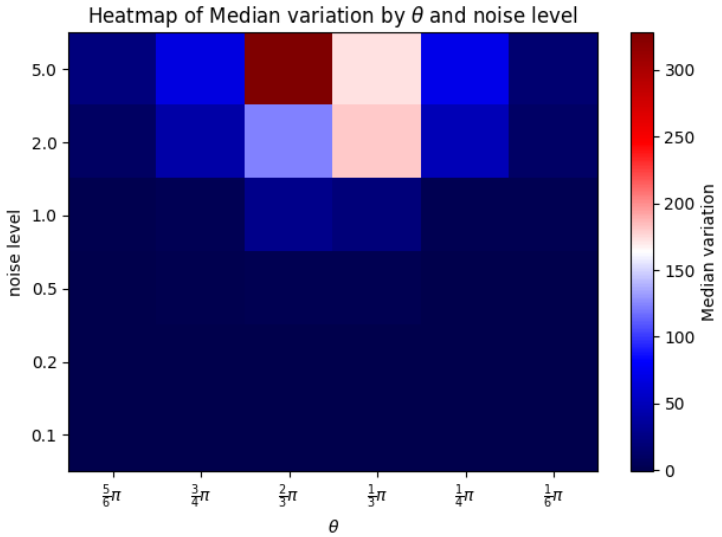


Figure 3: Median variation obtained varying the angle rotation θ and noise level σ.

Conclusion

We presents a robust approach to domain adaptation in linear regression using optimal transport, focusing on scenarios where domains are related by rotation. Our results demonstrate that the proposed method achieves better results in $\mathbb{R}^2$, consistently achieving better median performance compared to standard regression methods, even when the target domain size is limited. This highlights the method's potential for practical applications where training data is abundant, but data in the target domain is scarce. Several directions for future improvement and generalization emerge from this study:

1. **Understanding the origin of outliers and improving robustness:** The proposed method exhibits strong median performance but suffers from significant outliers. A deeper analysis of these outliers will be critical to designing a more robust algorithm that can handle such discrepancies more effectively.
2. **Generalization to Higher Dimensions:** The extension of the method from $\mathbb{R}^2$ to $\mathbb{R}^n$ using PCA for dimensionality reduction is promising. However, additional research is required to optimize this process.
3. **Incorporating Translations:** Expanding the framework to incorporate translations will significantly enhance its versatility.

These extensions, combined with the demonstrated effectiveness of our method in $\mathbb{R}^2$, provide a foundation for future work aimed at addressing more complex domain adaptation challenges in higher dimensions and diverse scenarios.

Bibliography

Monge G. (1781), Mémoire sur la théorie des déblais et des remblais.

Kantorovich L. (1942), On the transfer of masses (in russian), *Dokl. Akad. Nauk SSSR*, , 37, No. 7–8, pp. 227–229.

Courty N., Flamary R., Tuia D. and Rakotomamonjy A. (2016), Optimal Transport for Domain Adaptation, *IEEE Transactions on pattern analysis and machine intelligence*.

Arun K. S. , Huang T. S. and Blostein S. D. (1987), Least-squares fitting of two 3-D point sets.

Peyré G. and Cuturi M. (2019), Computational Optimal Transport, *Foundations and Trends in Machine Learning*.

Aligning Embeddings and Geometric Random Graphs: Informational Results and Computational Approaches for the Procrustes-Wasserstein Problem

Mathieu Even[1] & Luca Ganassali[2] & Jakob Maier[1] & Laurent Massoulié[1]

[1] *Inria Paris*
[2] *Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay*
luca.ganassali@universite-paris-saclay.fr

Résumé. Le problème de Procrustes-Wasserstein consiste à apparier deux nuages de points en dimension d, de façon non supervisée. Cette question possède de nombreuses applications en machine learning, notamment en traitement naturel du langage, et en computer vision. Nous formalisons ce problème comme suit : nous considérons deux nuages de points X, Y comprenant chacun n points dans $\mathbb{R}^d$, où Y est une version bruitée de X, à transformation orthogonale et permutation des points près. Cette description est fortement reliée à celle de l'alignement de graphes géométriques. Dans ce travail, nous envisageons le coût euclidien de transport entre les nuages de points comme mesure de performance de l'etimation. Nous établissons des résultats informationnel, dans les régimes de faible ($d \ll \log n$) et grande ($d \gg \log n$) dimension. Nous étudions également les aspects computationnels, en proposant l'algorithme Ping-Pong qui alterne entre l'estimation de la transformation orthogonale et celle de la permutation, initialisé par une relaxation convexe de Franke-Wolfe. Nous donnons des conditions suffisantes pour que cette méthode retrouve le signal. Enfin, nous comparons la performance de cette méthode aux approches précédemment suggérées, notamment par Grave et al. (2019).

Mots-clés. Inférence statistique, statistiques en grandes dimensions, graphes, transport optimal

Abstract. The Procrustes-Wasserstein problem consists in matching two high-dimensional point clouds in an unsupervised setting, and has many applications in natural language processing and computer vision. We consider a planted model with two datasets X, Y that consist of n datapoints in $\mathbb{R}^d$, where Y is a noisy version of X, up to an orthogonal transformation and a relabeling of the data points. This setting is related to the graph alignment problem in geometric models. In this work, we focus on the euclidean transport cost between the point clouds as a measure of performance for the alignment. We first establish information-theoretic results, in the high ($d \ll \log n$) and low ($d \gg \log n$) dimensional regimes. We then study computational aspects and propose the 'Ping-Pong algorithm', alternatively estimating the orthogonal transformation and the relabeling, initialized via a Franke-Wolfe convex relaxation. We give sufficient conditions for the method to retrieve the planted signal after one single step. We provide experimental results to compare the proposed approach with the state-of-the-art method of Grave et al. (2019).

Keywords. statistical inference, high dimensional statistics, graphs, optimal transport

Note of the authors: this extended abstract is a part of the introduction of our paper (Even et al. [2024]) presented at Neurips 2024, which can be found here.

1 Context

Finding an alignment between high dimensional vectors or across two point clouds of embeddings has been the focus of recent threads of research and has a variety of applications in computer vision, such as inferring scene geometry and camera motion from a stream of images [Tomasi and Kanade, 1992], as well as in natural language processing such as automatic unsupervised translation [Rapp, 1995, Fung, 1995].

Many practical algorithms proposed for this task view this problem as minimizing the distance across distributions in $\mathbb{R}^d$. Some approaches are based e.g. on optimal transport and Gromov-Wasserstein distance [Alvarez-Melis and Jaakkola, 2018] or adversarial learning [Zhang et al., 2017, Conneau et al., 2018]. Another line of methods adapt the iterative closest points procedure (ICP) – originally introduced in Besl and McKay [1992] for 3-D shapes – to higher dimensions [Hoshen and Wolf, 2018]. Another recent contribution is that of Grave et al. [2019], where a method is proposed to jointly learn an orthogonal transformation and an alignment between two point clouds by alternating the objectives in the corresponding minimization problem.

To formalize this problem, we consider a Gaussian model in which both datasets $X, Y \in \mathbb{R}^{d\times n}$ (or two point clouds of n datapoints in $\mathbb{R}^d$) are sampled as follows. First, $X = (x_1, \dots, x_n)$ is a collection of i.i.d. $\mathcal{N}(0, I_d)$ Gaussian vectors, and $Y = (y_1, \dots, y_n)$ is a noisy version of $X = (x_1, \dots, x_n)$, up to an orthogonal transformation $Q^\star$ and a relabeling $\pi^\star : [n] \to [n]$ of the data points, that is:

$$\forall i \in [n]\,, \quad y_i = Q^\star x_{\pi^\star(i)} + \sigma z_i\,, \tag{1}$$

or, in matrix form:

$$Y = Q^\star X (P^\star)^\top + \sigma Z\,,$$

where $Z = (z_1, \dots, z_n) \in \mathbb{R}^{d\times n}$ is also made of i.i.d. $\mathcal{N}(0, I_d)$ Gaussian vectors, $P^\star$ is the permutation matrix associated with some permutation $\pi^\star$, and $\sigma > 0$ is the noise parameter. Recovering (in some sense that will be made precise in the sequel) the (unknown) permutation $\pi^\star$ and orthogonal transformation $Q^\star$ defines the *Procrustes-Wasserstein problem* (sometimes abbreviated as PW in the sequel), which will be the focus of this study.

The practical approaches previously mentioned have shown good empirical results and are often scalable to large datasets. However, they suffer from a lack of theoretical results to guarantee their performance or to exhibit regimes where they fail. Model (1) described here above appears to be the simplest one to obtain such guarantees. We are interested in pinning down the *fundamental* limits of the Procrustes-Wasserstein problem, hence providing an ideal baseline for any computational method to be compared to, before delving into computational aspects. Our contributions are as follows:

(*i*) We define a planted model for the Procrustes-Wassertein problem and discuss the appropriate choice of metrics to measure the performance of any estimator. Based on these metrics, we establish[1]:

(*i.a*) information-theoretic results in the high-dimensional $d \gg \log n$ regime which was not explored before for this problem;

(*i.b*) new information-theoretic results in the low-dimensional regime ($d \ll \log n$) for our metric of performance (the L^2 transport cost), which substantially differ from those obtained in Wang et al. [2022] for the overlap.

(*ii*) We study computational aspects and propose the 'Ping-Pong algorithm', alternatively estimating the orthogonal transformation and the relabeling, initialized via a Franke-Wolfe convex relaxation. This method is quite close to that proposed in Grave et al. [2019] although the alternating part differs. We give sufficient conditions for the method to retrieve the planted signal after one single step.

(*iii*) Finally, we provide experimental results to compare the proposed approach with the state-of-the-art method of Grave et al. [2019].

2 Discussion and related work

One can check that under the above model (1), the maximum likelihood (ML) estimators of $(P^\star, Q^\star)$ given (X, Y) is given by:

$$(\hat{P}, \hat{Q}) \in \operatorname*{arg\,min}_{(P,Q)\in\mathcal{S}_n\times\mathcal{O}(d)} \frac{1}{n}\|XP^\top - Q^\top Y\|_F^2 = \operatorname*{arg\,min}_{(P,Q)\in\mathcal{S}_n\times\mathcal{O}(d)} \|QX - YP\|_F^2\,, \tag{2}$$

which is strictly equivalent[2] to the formulation of the non-planted problem of Grave et al. [2019]. Exactly solving the joint optimization problem (2) is non convex and difficult in general. However, if $P^\star$ is known then (2) boils down to the following *orthogonal Procrustes problem*:

$$\hat{Q} \in \operatorname*{arg\,min}_{Q\in\mathcal{O}(d)} \frac{1}{n}\|XP^\star - Q^\top Y\|_F^2\,, \tag{3}$$

which has a simple closed form solution given by $\hat{Q} = UV^\top$ where $USV^\top$ is the singular value decomposition (SVD) of $Y(XP^\star)^\top$ (see Schönemann [1966]). Conversely, when $Q^\star$ is known, (2) amounts to the following *linear assignment problem* (LAP in the sequel):

$$\operatorname*{arg\,min}_{P\in\mathcal{S}_n} \frac{1}{n}\|XP^\top - Q^\star Y\|_F^2 = \operatorname*{arg\,max}_{P\in\mathcal{S}_n} \frac{1}{n}\langle XP^\top, Q^\star Y\rangle, \tag{4}$$

which can be solved in polynomial time, e.g. in cubic time by the celebrated Hungarian algorithm [Kuhn, 1955], or more efficiently at the price of regularizing the objective and using the celebrated Sinkhorn algorithm [Cuturi, 2013].

[1] this very dichotomy ($d \gg \log n$ versus $d \ll \log n$) is fundamental in high dimensional statistics and not a mere artifact of our rationale; see Remark 1 of the full paper (Even et al. [2024]).

[2] their matrices X and Y are the transposed versions of ours.

Previous results when $Q^\star$ is known. As seen above, when $Q^\star$ is known (assume e.g. $Q^\star = I_d$), the Procrustes-Wasserstein problem reduces to a simpler objective, that of aligning Gaussian databases. This problem has been studied by Dai et al. [2019, 2023] in the context of feature matching. Kunisky and Niles-Weed [2022] study the same problem as a geometric extension of planted matching and establish state-of-the-art statistical bounds in the Gaussian model in the low-dimensional ($d \ll \log n$), logarithmic ($d \sim a \log n$) and high-dimensional ($d \gg \log n$) regimes. In particular, they show that exact recovery is feasible in the logarithmic regime $d \sim a \log n$ if $\sigma^2 < \frac{1}{e^{4/a}-1}$, and in the high-dimensional regime if $\sigma^2 < (1/4 - \varepsilon)\frac{d}{\log n}$. Note that in this problem, there is no computational/statistical gap since the LAP is always solvable in polynomial time.

Geometric graph alignment. Strongly connected to the Procrustes-Wassertein problem is the topic of graph alignment where the instances come from a geometric model. Wang et al. [2022] investigate this problem for complete weighted graphs. In their setting, given a permutation $\pi^\star$ on $[n]$ and n i.i.d. pairs of correlated Gaussian vectors $(X_{\pi^\star(i)}, Y_i)$ in $\mathbb{R}^d$ with noise parameter σ, they observe matrices $A = X^\top X$ and $B = Y^\top Y$ (i.e all inner products $\langle X_i, X_j \rangle$ and $\langle Y_i, Y_j \rangle$) and are interested in recovering the hidden vertex correspondence $\pi^\star$. The maximum likelihood estimator in this setting writes

$$\operatorname*{arg\,min}_{P \in \mathcal{S}_n} \frac{1}{n} \|X^\top X P - P Y^\top Y\|_F^2 = \operatorname*{arg\,max}_{P \in \mathcal{S}_n} \frac{1}{n} \langle P^\top X^\top X P, Y^\top Y \rangle \,, \tag{5}$$

which is an instance of the *quadratic assignment problem* (QAP in the sequel), known to be NP-hard in general, as well as some of its approximations [Makarychev et al., 2014]. In fact, we have the following informal equivalence[3]:

Lemma 1 (Informal)**.** *PW and geometric graph alignement are equivalent, that is, one knows how to (approximately) solve the former iff they know how to (approximately) solve the latter.*

Wang et al. [2022] focus on the low-dimensional regime $d = o(\log n)$, where geometry plays the most important role. They prove that exact (resp. almost exact) recovery of $\pi^\star$ is information-theoretically possible as soon as $\sigma = o(n^{-2/d})$ (resp. $\sigma = o(n^{-1/d})$). They conduct numerical experiments which suggest good performance of the celebrated Umeyama algorithm [Umeyama, 1988], which is confirmed by a follow-up work by Gong and Li [2024] analyzing the Umeyama algorithm (which is polynomial time in the low dimensional regime $d = o(\log n)$) in the same setting and shows that it achieves exact (resp. almost exact) recovery of $\pi^\star$ if $\sigma = o(d^{-3} n^{-2/d})$ (resp. $\sigma = o(d^{-3} n^{-1/d})$), hence coinciding with the information thresholds up to a poly(d) factor. However, their algorithm is of time complexity at least $\Omega(2^d n^3)$, which is not polynomial in d. This is why we do not include this method in our baselines.

We emphasize that our results clearly depart from those obtained in Wang et al. [2022] and Gong and Li [2024], because (i) we are also interested in the high dimensional case $d \gg \log n$, and (ii) we work with a different performance metric which provides less stringent conditions for the recovery to be feasible. A summary of previous informational results together with ours is given in Table 1.

[3] see Appendix A of the full paper Even et al. [2024] for a proof.

Reference	Setting	Metrics	Regime	Condition
Kunisky and Niles-Weed [2022]	$Q^\star = I_d$	ov for $P^\star$	$d \ll \log n$	$\sigma \ll n^{-2/d}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = 0$ $\sigma \ll n^{-1/d}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = o(1)$
			$d \sim a \log n$	$\sigma < (e^{4/a} - 1)^{-1/2}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = 0$ $\sigma < ((2e^{1/a} - 1)^2 - 1)^{-1/2}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = o(1)$
			$d \gg \log n$	$\sigma < (1/2 - \varepsilon)(d/\log n)^{1/2}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = 0$
Wang et al. [2022]	$A = X^\top X, B = Y^\top Y$	ov for $P^\star$	$d \ll \log n$	$\sigma \ll n^{-2/d}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = 0$ $\sigma \ll n^{-1/d}$ for $\mathrm{ov}(\hat{\pi}, \pi^\star) = o(1)$
This paper	X, Y from (1)	c^2 for $P^\star$, ℓ^2 for $Q^\star$	$d \ll \log n$	$\sigma \ll d^{-1/2}$ for $c^2(\hat{\pi}, \pi^\star) = \ell^2(\hat{Q}, Q^\star) = o(d)$
			$d \gg \log n$	$\sigma \ll 1$ for $c^2(\hat{\pi}, \pi^\star) = \ell^2(\hat{Q}, Q^\star) = o(d)$

Table 1: Summary of previous informational results, together with the ones in this paper

On the orthogonal transformation $Q^\star$. Generalizing the standard linear assignment problem, our model described above in (1) introduces an additional orthogonal transformation $Q^\star$ across the datasets. This orthogonal transformation can be motivated in the context of aligning embeddings in a high-dimensional space: indeed, the task of learning embeddings is often agnostic to orientation in the latent space. In other words, two point clouds may represent the same data points while having different global orientations. Hence, across different data sets, learning this orientation shift is crucial in order to compare (or align) the point clouds. As an illustration of this fact, Xing et al. [2015] provides empirical evidence that orthogonal transformations are particularly adapted for bilingual word translation.

Discussing the method proposed in Grave et al. [2019]. We conclude this introduction by discussing the work of Grave et al. [2019]. Their proposed algorithm is as follows. At each iteration t, given a current estimate Q_t of the orthogonal transformation, we sample mini-batches X_t, Y_t of same size b and find the optimal matching P_t between $Y_t Q_t^\top$ and X_t, via solving a linear assignment problem of size b. This matching P_t in turn helps to refine the estimation of the orthogonal transformation via a projected gradient descent step, and the procedure repeats. This method has the main advantage to be scalable to very large datasets and to perform well in practice ; however, no guarantees are given for this method, and in particular the mini-batch step which can justifiably raise some concerns. Indeed, since $X_t = (x_{t,j})_{j \in [b]}$ and $Y_t = (y_{t,j})_{j \in [b]}$ are chosen independently, if $b \ll \sqrt{n}$ it is likely that for any matching π_t the pairs $(x_{t,j}, y_{t,\pi_t(j)})$ always correspond to disjoint pairs, and thus aligning $Y_t Q_t^\top$ and X_t does not reveal any useful information about the true $P^\star$ – this is even more striking when the data is non-isotropic.

References

David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1214. URL `https://aclanthology.org/D18-1214`.

P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. doi: 10.1109/34.121791.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Osman E. Dai, Daniel Cullina, and Negar Kiyavash. Database alignment with gaussian features. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3225–3233. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/dai19b.html.

Osman Emre Dai, Daniel Cullina, and Negar Kiyavash. Gaussian database alignment and gaussian planted matching, 2023.

Mathieu Even, Luca Ganassali, Jakob Maier, and Laurent Massoulié. Aligning embeddings and geometric random graphs: Informational results and computational approaches for the procrustes-wasserstein problem. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 70730–70764. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/829300b183943850930c3ce36fe04268-Paper-Conference.pdf.

Pascale Fung. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Third Workshop on Very Large Corpora*, 1995. URL https://aclanthology.org/W95-0114.

Shuyang Gong and Zhangsong Li. The umeyama algorithm for matching correlated gaussian geometric models in the low-dimensional regime, 2024.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR, 2019.

Yedid Hoshen and Lior Wolf. Non-adversarial unsupervised word translation, 2018.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. doi: https://doi.org/10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109.

Dmitriy Kunisky and Jonathan Niles-Weed. *Strong recovery of geometric planted matchings*, pages 834–876. 2022. doi: 10.1137/1.9781611977073.36. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611977073.36.

Konstantin Makarychev, Rajsekar Manokaran, and Maxim Sviridenko. Maximum quadratic assignment problem: Reduction from maximum label cover and lp-based approximation algorithm. *ACM Trans. Algorithms*, 10(4), aug 2014. ISSN 1549-6325. doi: 10.1145/2629672. URL https://doi.org/10.1145/2629672.

Reinhard Rapp. Identifying word translations in non-parallel texts. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322, Cambridge, Massachusetts, USA, June 1995. Association for Computational Linguistics. doi: 10.3115/981658.981709. URL https://aclanthology.org/P95-1050.

Peter Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. URL https://EconPapers.repec.org/RePEc:spr:psycho:v:31:y:1966:i:1:p:1-10.

Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, Nov 1992. ISSN 1573-1405. doi: 10.1007/BF00129684. URL https://doi.org/10.1007/BF00129684.

S. Umeyama. An eigendecomposition approach to weighted graph matching problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(5):695–703, 1988. doi: 10.1109/34.6778.

Haoyu Wang, Yihong Wu, Jiaming Xu, and Israel Yolou. Random graph matching in geometric models: the case of complete graphs. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3441–3488. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/wang22a.html.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1104. URL https://aclanthology.org/N15-1104.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1179. URL https://aclanthology.org/P17-1179.

Prédiction par krigeage de mesures de probabilité. Application en simulation numérique pour les études de sûreté nucléaire.

Florian Gossard[1], François Bachoc[1,2], Jean Baccou[3], Thibaut Le Gouic[4], Jacques Liandrat[4], Tony Glantz[3]

[1] *IMT, Univ. Paul Sabatier, F-31062 Toulouse, France, florian.gossard@math.univ-toulouse.fr, françois.bachoc@math.univ-toulouse.fr*
[2] *Institut Universitaire de France (IUF), France*
[3] *ASNR, France, jean.baccou@asnr.fr, tony.glantz@asnr.fr*
[4] *Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France, thibaut.le-gouic@centrale-marseille.fr, jacques.liandrat@centrale-marseille.fr*

Résumé. La prédiction de données complexes (i.e. non scalaires) ressort dans plusieurs applications en sureté nucléaire notamment en thermohydraulique pour les études en situation accidentelle par simulation numérique. En pratique la quantité de données simulées peut être insuffisante pour analyser finement un phénomène d'intérêt en raison du coût de calcul d'une simulation et on a recours à l'utilisation de modèles rapides à évaluer (métamodèles). Nous présentons ici la construction d'un métamodèle à base de krigeage pour des données complexes de type mesure de probabilité. Pour cela, nous exploitons la distance de Wasserstein issue du transport optimal et proposons de construire l'estimateur comme un barycentre. Nous appliquons ensuite cette approche pour la prédiction de mesures associées à des températures simulées dans le coeur d'un réacteur nucléaire lors d'un accident de perte de réfrigérant.

Mots-clés. Krigeage, Mesure de probabilités, distance de Wasserstein, Sûreté nucléaire

Abstract. The prediction of complex data is a key topic in several applications related to nuclear safety, particularly in thermohydraulics for the analysis of accidental scenarios by numerical simulation. In practice, the amount of simulated data may be insufficient to perform a detailed analysis of a phenomenon of interest due to the high computational cost of a simulation. To circumvent this difficulty, fast-to-evaluate models or metamodels can be used. In this paper we present the construction of a Kriging-based metamodel for complex data in the form of probability measures. We exploit the Wasserstein distance derived from optimal transport theory and propose to construct the Kriging estimator as a Wasserstein barycenter. This approach is then applied to the prediction of measures associated with the temperatures simulated in the core of a nuclear reactor during a loss of primary coolant accident.

Keywords. Kriging, Probability measure, Wasserstein distance, Nuclear safety

1 Introduction

L'étude de systèmes industriels complexes tels qu'un réacteur nucléaire nécessite d'avoir recours à des codes de calcul permettant de simuler numériquement plusieurs phénomènes d'intérêt. En pratique, la quantité de données simulées peut être insuffisante pour analyser précisément un phénomène sur tout son domaine d'intérêt en raison du coût numérique d'une simulation. On s'oriente alors vers la construction d'un métamodèle permettant d'approcher à faible coût la grandeur d'intérêt simulée dans des situations différentes de celles déjà simulées. Parmi les approches classiques de métamodélisation, le krigeage [Rasmussen and Williams, 2006] est largement utilisé dans le domaine de l'exploration des codes de calcul complexes avec un nombre réduit de données d'apprentissage. Toutefois, dans les applications, le Krigeage se focalise souvent sur des grandeurs d'intérêt de type scalaire (i.e. la simulation fournit une valeur). L'objectif de ce travail[1] est de tester plusieurs approches à base de krigeage pour traiter des sorties d'intérêt qui sont des mesures de probabilité discrètes décrivant par exemple la répartition de températures simulées dans le cœur d'un réacteur en situation accidentelle.

Dans cet article, nous rappelons dans la Section 2 les éléments principaux de la construction de l'estimateur du krigeage dans le cas scalaire puis décrivons une extension pour des mesures de probabilité. La Section 3 fournit ensuite une application de ces estimateurs dans le cadre des études en sûreté nucléaire à l'ASNR (Autorité de Sûreté Nucléaire et de Radioprotection).

2 Méthodes de krigeage

2.1 Grandeur d'intérêt de type scalaire

Dans ce cas, le code de calcul peut formellement s'écrire comme une fonction de transfert inconnue $f : \mathcal{D} \subset \mathbb{R}^d \mapsto \mathbb{R}$. Les simulations disponibles sont alors des évaluations de f pour un ensemble de combinaisons de valeurs de ses variables d'entrée.

Nous rappelons brièvement dans la suite les étapes de la c onstruction d'un estimateur du krigeage en adoptant la vision géostatistique [Cressie, 1993] car c'est celle qui facilitera l'extension du krigeage au cas de mesures de probabilités.
Les observations de f sont considérées comme des réalisations d'un ensemble de variables aléatoires provenant d'un processus aléatoire $\mathcal{F}(x)$ qui se décompose comme suit: $\forall x \in \mathcal{D},\ \mathcal{F}(x) = \mu(x) + \delta(x)$ où $\mu(x)$ est la partie déterministe du processus et $\delta(x)$ est un processus aléatoire de moyenne nulle et spatialement corrélé. Sous les hypothèses μ constante, stationarité et isotropie[2] du processus, la corrélation spatiale s'obtient en estimant le semi-variogramme $\gamma(x, x') = \frac{1}{2}\mathbf{E}\left(\mathcal{F}(x) - \mathcal{F}(x')\right)^2$ à partir de l'ensemble des observations noté

[1] Réalisé dans le cadre du projet QUTHY du programme NEEDS (accord de coopération géré par le CNRS dans le domaine du nucléaire entre CNRS, Andra, BRGM, CEA, EDF, Framatome, IRSN et Orano).

[2] On peut s'affranchir de ces deux hypothèses mais elles sont retenues ici pour simplifier la description de la construction de la méthode.

$\{f(x_i)\}_{i=0,...,N-1}$ par le semi-variogramme expérimental

$$\gamma_{exp}(h) = \frac{1}{2Card(N(h))} \sum_{(m,n)\in N(h)} (f(x_m) - f(x_n))^2$$

avec $N(h) = \{(m,n) \in \{0,...,N-1\}^2,\ h - \epsilon \leq ||x_m - x_n||_2 \leq h + \epsilon\}$ (ϵ dépend du problème et peut être fixé par l'utilisateur) et pour chaque h tel que $Card(N(h))$, le cardinal de $N(h)$, est suffisamment grand. Les candidats pour les modèles d'ajustement du semi-variogramme expérimental sont à choisir parmi une famille de modèles de semi-variogrammes valides.

L'estimateur du krigeage ordinaire du processus aléatoire $\mathcal{F}$ au point $x^\star$, noté $\mathcal{P}_{ko}(\mathcal{F}, x^\star)$, est le meilleur prédicteur linéaire sans biais qui minimise la variance d'estimation. Il s'écrit:

$$\mathcal{P}_{ko}(\mathcal{F}, x^\star) = \sum_{i=0}^{N-1} \bar{\lambda}_i \mathcal{F}(x_i), \tag{1}$$

où

$$(\bar{\lambda}_i)_{i=0,...,N-1} = \text{argmin} \left\{ \mathbf{E}\left(\mathcal{F}(x^\star) - \mathcal{P}_{ko}(\mathcal{F}, x^\star)\right)^2, \sum_{i=0}^{N-1} \lambda_i = 1 \right\}. \tag{2}$$

En introduisant $u = \left(\bar{\lambda}_0, \ldots, \bar{\lambda}_{N-1}\right)^T$ et $\mathcal{L}$ un multiplicateur de Lagrange pour prendre en compte la condition de non biais, on peut montrer que $(u, \mathcal{L})^T$ est solution d'un système matriciel faisant intervenir le semi-variogramme [Cressie, 1993].

Dans le domaine de l'exploration des codes de calcul [Sacks et al., 1989], on suppose en plus que le processus aléatoire $\mathcal{F}$ est gaussien et on estime les paramètres de sa covariance (au lieu du semi-variogramme) par une méthode de maximum de vraisemblance. Toutefois, le système linéaire à résoudre pour obtenir les poids du krigeage est similaire à celui faisant intervenir le semi-variogramme.

2.2 Grandeur d'intérêt de type mesure de probabilité

Dans ce cas, on interprète le code de calcul comme une fonction de transfert inconnue $f : \mathcal{D} \subset \mathbb{R}^d \mapsto \mathcal{H}$ où $\mathcal{H}$ est l'espace $\mathcal{P}_2(\mathbb{R}^d)$ des mesures de probabilité μ telles que $\int \|x\|^2 d\mu(x) < \infty$. Pour construire un estimateur du krigeage prenant en compte la spécificité de $\mathcal{H}$, on peut s'orienter vers des approches issues du domaine des statistiques orientées objet [Menafoglio and Secchi, 2017]. Ces dernières consistent à interpréter chaque donnée comme un point dans un espace caractéristique et à adapter les méthodes de krigeage opérant sur des grandeurs scalaires dans ce nouvel espace.

En remarquant que l'estimateur linéaire (1) peut s'interpréter comme un barycentre i.e. $\mathcal{P}_{ko}(\mathcal{F}, x^\star) = \text{argmin}_{y\in\mathcal{H}} \left\{ \sum_{i=0}^{N-1} \bar{\lambda}_i \left(\mathcal{F}(x_i) - y\right)^2 \right\}$, on peut proposer formellement un estimateur du krigeage pour un espace $\mathcal{H}$ muni d'une distance $d_{\mathcal{H}}$:

$$\mathcal{P}_{ko}(\mathcal{F}, x^\star) = \text{argmin}_{y\in\mathcal{H}} \left\{ \sum_{i=0}^{N-1} \bar{\lambda}_i d^2_{\mathcal{H}} \left(\mathcal{F}(x_i), y\right) \right\}, \tag{3}$$

où

$$(\bar{\lambda}_i)_{i=0,\dots,N-1} = \text{argmin} \left\{ \mathbf{E}\left(d^2_{\mathcal{H}} \left(\mathcal{F}(x^\star), \mathcal{P}_{ko}(\mathcal{F}, x^\star)\right)\right), \sum_{i=0}^{N-1} \lambda_i = 1 \right\}. \tag{4}$$

Une distance pertinente pour comparer deux mesures de probabilité μ et ν est la distance de Wasserstein d'ordre 2 provenant du domaine du transport optimal [Santambrogio, 2015]. Elle s'écrit $W_2(\mu,\nu) = \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times X} \|x-y\|_2^2 \,\mathrm{d}\gamma(x,y)\right)^{\frac{1}{2}}$ où $\Gamma(\mu,\nu)$ est l'ensemble des couplages entre les deux mesures μ et ν. Ce sera la distance $d_{\mathcal{H}}$ utilisée dans les expressions (3) et (4). Dans le cas général de mesures à support dans $\mathbb{R}^d$, le problème de calcul des poids (4) est difficile à résoudre car l'estimateur du krigeage n'est plus linéaire [Gouet et al., 2015]. Toutefois, pour des mesures à support dans $\mathbb{R}$, on peut se ramener à un estimateur linéaire sur les fonctions quantiles [Balzanella and Irpino, 2020].

3 Application au cas du renoyage

3.1 Cas d'étude

On s'intéresse ici à la simulation numérique d'un accident de perte de réfrigérant primaire dans un réacteur à eau pressurisée et plus particulièrement à l'étude du phénomène de renoyage du coeur du réacteur à l'aide du logiciel DRACCAR [Bascou et al., 2015]. Le cas test est une simulation dans le temps de la carte des températures de quelques crayons du coeur du réacteur (100 points de discrétisation) en fonction de la configuration du système. Cette configuration est caractérisée par 6 paramètres physiques (Table 1). 101 simulations ont été réalisées en faisant varier ces 6 paramètres .

	min	max	mean
dtsub (K)	0.217	59.578	27.655
pout ($\times 10^6$) (Pa)	0.222	5.968	3.135
vin (m/s)	0.016	0.079	0.042
p1rod (W)	612.542	3440.880	2077.480
pshroud (W)	344.640	24407.170	12738.210
tclad0 (K)	730.183	971.501	847.148

Table 1: Intervalle de variation des 6 paramètres physiques.

Pour chaque simulation, on se limite à la carte obtenue à l'instant où la température maximale est atteinte. Il est difficile d'explorer précisément l'ensemble des configurations du domaine physique en raison du temps de calcul de DRACCAR qui peut aller de quelques minutes à un mois suivant le domaine simulé (nombre de crayons considérés) et le transitoire considéré. Les approches de métamodélisation apportent alors une solution pour contourner cette limitation. En pratique, on s'intéresse à des caractéristiques statistiques de la carte

telles que la moyenne, la variance ou un quantile d'intérêt. Plutôt que de faire un métamodèle par caractéristique, nous proposons de faire un krigeage opérant directement sur la mesure discrète construite à partir des valeurs de la température de la carte puis d'en déduire les caractéristiques d'intérêt. Pour cela, on s'appuie sur l'extension du krigeage classique décrit dans la Section 2.2.

3.2 Résultats numériques

Nous évaluons les performances de l'approche de krigeage de mesures de probabilité et les comparons avec celles de krigeages plus classiques permettant de prédire la carte des températures. Plus précisément, quatres modèles sont testés:

Modèle 1 : en partant des cartes discrétisées en 100 points, une réduction de dimension par analyse en composantes principales (ACP) est d'abord effectuée puis un modèle de krigeage est construit pour chacune des trois premières composantes principales qui portent 99.9% de la variance. Un modèle de semi-variogramme isotrope est supposé avec une détermination de ses paramètres par la méthode des moindres carrés.

Modèle 2 : même approche que précédemment, la seule différence est que l'on modélise chaque carte comme des réalisations d'un processus gaussien, ce qui permet d'estimer les paramètres de sa covariance par maximum de vraisemblance.

Modèle 3 : en partant des valeurs de la température, on applique le krigeage de mesures de probabilité en supposant un semi-variogramme isotrope dont les paramètres sont estimés par la méthode des moindres carrés.

Modèle 4 : même approche que précédemment, la seule différence est que l'on rajoute dans le problème de krigeage une contrainte de non-négativité des poids en s'appuyant sur l'approche empririque suggérée dans [Barnes and Johnson, 1984].

Les performances de chaque modèle (Figure 1) sont évaluées par le calcul de la RMSE (Root Mean Square Error) associée à la moyenne, à la variance et au quantile à 95%. De plus, une extension de ce critère de performance où l'écart quadratique est remplacé par le carré de la distance de Wasserstein est utilisé pour comparer les mesures prédites. Pour prendre en compte l'effet des données, 90 couples d'ensembles (apprentissage,test) sont générés. Le modèle 3 présente les meilleures performances notamment pour la prédiction du quantile à 95%, ce qui est intéressant en pratique car les valeurs élevées de la température conduisent à des dégradations plus importantes du système. On note également que le modèle 4 (contrainte de non négativité des poids) est le moins performant.

4 Conclusion

Un estimateur pour le krigeage de mesures de probabilité a été proposé en s'appuyant sur une extension du krigeage classique via la construction du barycentre de Wasserstein. Son intérêt est qu'il permet en un seul krigeage de prédire des caractéristiques statistiques telle que la moyenne, la variance ou des quantiles. L'application en thermohydraulique dans le cadre de la sûreté nucléaire a montré que cet estimateur conduit à une plus grande précision

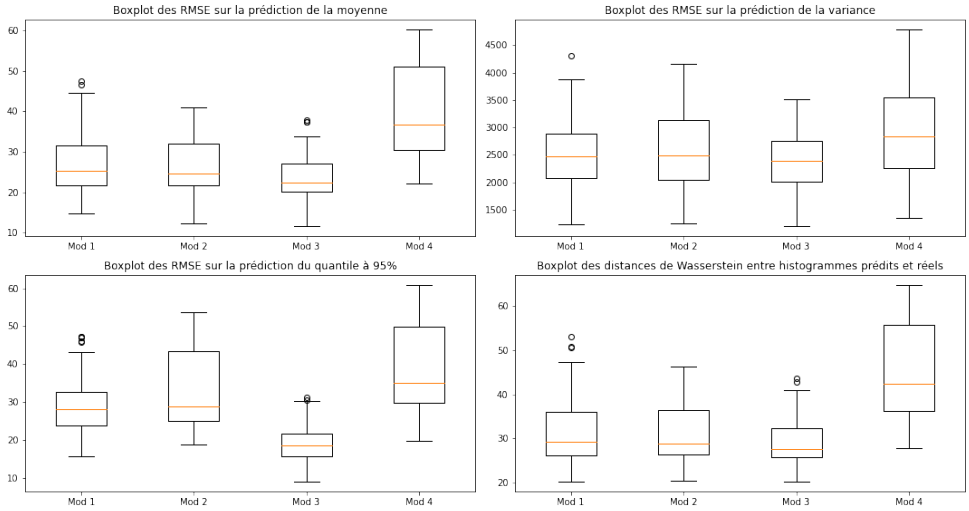


Figure 1: Boxplots des critères de performance pour les 4 modèles.

notamment pour des quantiles élevés qui sont des grandeurs d'intérêt pour la sûreté. Pour la suite, on souhaiterait améliorer, pour les modèles concernés, l'estimation des paramètres du variogramme en les optimisant par validation croisée [Bachoc, 2013]. Un objectif à plus long terme est de développer des modèles prenant en compte la dimension spatiale des données.

Bibliographie

Bachoc, F. (2013), Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification, *Computational Statistics & Data Analysis*, 66, pp. 55-69.

Balzanella, A., Irpino, A. (2020), Spatial Prediction and Spatial Dependence Monitoring on Georeferenced Data Streams, *Statistical Methods & Applications*, 29(1), pp. 101-128.

Barnes, R., Johnson, T. (1984), Positive kriging, *Verly, G., ed., Geostatistics for Natural Resources Characterization: v. 1, Reidel, Dordrecht, Holland*,pp l-244.

Bascou, S., De Luze, O., Ederli, S., Guillard, G. (2015), Development and validation of the multi-physics DRACCAR code, *Annals of Nuclear Energy*, 84, pp. 1-18.

Cressie, N. (1993), *Statistics for Spatial Data*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, J. Wiley.

Gouet, R., Lopez, A., Ortiz, J.M. (2015), Geodesic Kriging in the Wasserstein Space, *Proceedings of the 17th Annual Conference of the International Association for Mathematical Geosciences IAMG*.

Menafoglio, A.,Secchi, P. (2017), Statistical Analysis of Complex and Spatially Dependent Data: A Review of Object Oriented Spatial Statistics, *European journal of operational research*, 258(2), pp. 401-410.

Rasmussen, C.E., Williams, C.K.I. (2006), *Gaussian Processes for Machine Learning*, the MIT Press.

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P. (1989), Design and Analysis of Computer Experiments, *Statist. Sci.*, 4(4): 409-423.

Santambrogio, F. (2015), *Optimal Transport for Applied Mathematicians*, Birkhauser Verlag AG.

Barycentres de Mesures pour Coûts de Transport Génériques

Eloi Tanguy[1] & Julie Delon[1] & Nathaël Gozlan[1]

[1] *Université Paris Cité, CNRS, MAP5, F-75006 Paris, France*

Résumé. Les barycentres de Wasserstein représentent des distributions moyennes entre plusieurs mesures de probabilité pour la distance de Wasserstein. Le calcul numérique de ces barycentres est notoirement difficile. Une approche courante consiste à utiliser les itérations de Sinkhorn, où un terme de régularisation entropique est introduit afin de rendre le problème plus tractable. Une autre approche repose sur l'utilisation de méthodes de point fixe, similaires à celles employées pour le calcul des moyennes de Fréchet sur les variétés. La convergence de telles méthodes pour les barycentres 2-Wasserstein, en particulier avec une fonction de coût quadratique et des mesures absolument continues, a été étudiée par Alvarez-Esteban et al. dans [1]. Nous explorons les principales idées sous-jacentes à cette méthode de point fixe et examinons comment elle peut être généralisée pour prendre en compte des coûts de transport plus variés et des mesures de probabilité génériques, étendant ainsi son applicabilité à un éventail plus large de problèmes. Nous établissons des résultats de convergence pour cette approche et illustrons son comportement numérique sur plusieurs problèmes de barycentre.

Mots-clés. Transport Optimal, Barycentre, Méthode de Point-Fixe...

Abstract. Wasserstein barycentres represent average distributions between multiple probability measures for the Wasserstein distance. The numerical computation of Wasserstein barycentres is notoriously challenging. A common approach is to use Sinkhorn iterations, where an entropic regularisation term is introduced to render the problem more tractable. Another approach involves using fixed-point methods, akin to those employed for computing Fréchet means on manifolds. The convergence of such methods for 2-Wasserstein barycentres, specifically with a quadratic cost function and absolutely continuous measures, was studied by Alvarez-Esteban et al. in [1]. We explore the main ideas behind this fixed-point method and explore how it can be generalised to accommodate more diverse transport costs and any probability measures, thereby extending its applicability to a broader range of problems. We show convergence results for this approach and illustrate its numerical behaviour on several barycentre problems.

Keywords. Optimal Transport, Barycenter, Fixed-Point Method...

Note: this presentation summarises a recent work of the authors [6].

Problem Formulation

We work with probability measures ν_k on compact metric spaces $(\mathcal{Y}_k, d_{\mathcal{Y}_k})_{k\in[\![1,K]\!]}$, of which we will seek a barycentre μ in a compact metric space $(\mathcal{X}, d_{\mathcal{X}})$. To compare a measure $\nu_k \in \mathcal{P}(\mathcal{Y}_k)$ and another measure $\mu \in \mathcal{P}(\mathcal{X})$ we consider continuous cost functions $c_k : \mathcal{X} \times \mathcal{Y}_k \longrightarrow \mathbb{R}_+$. A barycentre will be a minimiser of the sum of the transport costs with respect to the measure ν_k, which leads to the following energy functional for a measure $\mu \in \mathcal{P}(\mathcal{X})$:

$$V(\mu) := \sum_{k=1}^{K} \mathcal{T}_{c_k}(\mu, \nu_k), \tag{1}$$

hence our minimisation problem reads

$$\operatorname*{argmin}_{\mu\in\mathcal{P}(\mathcal{X})} V(\mu). \tag{2}$$

The definition of a barycentre between measures ν_k can be seen as a lifting of a notion of barycentre within $\mathcal{X}$ of points $(y_1, \cdots, y_K) \in \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_K$. To formalise this intuition, we will make the following assumption throughout this note:

Assumption 1. *For all $(y_1, \cdots, y_K) \in \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_K$, the set $\operatorname*{argmin}_{x\in\mathcal{X}} \sum_{k=1}^{K} c_k(x, y_k)$ has a unique element.*

The uniqueness of the optimisation problem in Assumption 1 allows us to introduce the ground barycentre function B:

$$B : \left\{ \begin{array}{ccc} \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_K & \longrightarrow & \mathcal{X} \\ (y_1, \cdots, y_K) & \longmapsto & \operatorname*{argmin}_{x\in\mathcal{X}} \sum_{k=1}^{K} c_k(x, y_k). \end{array} \right. \tag{3}$$

Defining a Fixed-Point Method

We define a sequence $(\mu_t) \in \mathcal{P}(\mathcal{X})^{\mathbb{N}}$ that will approach a barycentre of fixed measures $\nu_k \in \mathcal{P}(\mathcal{Y}_k)$. We propose a modified version of the iterated scheme from [1] to solve Eq. (2). To define an iteration mapping, for $\mu \in \mathcal{P}(\mathcal{X})$, we consider the set of multi-marginal couplings

$$\Gamma(\mu) := \left\{ \gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_K) \ : \ \forall k \in [\![1, K]\!], \ \gamma_{0,k} \in \Pi^*_{c_k}(\mu, \nu_k) \right\}, \tag{4}$$

where, for all k, $\gamma_{0,k}$ denotes the $\mathcal{X} \times \mathcal{Y}_k$ marginal of γ and $\Pi^*_{c_k}(\mu, \nu_k)$ denotes the set of all optimal couplings for the transport problem between μ and ν_k associated to the cost function c_k. The existence of such multi-couplings is a consequence of the well-known "gluing lemma" (see [5] Lemma 5.5). The following multi-coupling provides an explicit element of $\Gamma(\mu)$ given $\pi_k \in \Pi^*_{c_k}(\mu, \nu_k)$:

$$\gamma(\mathrm{d}x, \mathrm{d}y_1, \cdots, \mathrm{d}y_K) := \mu(\mathrm{d}x)\pi_1^x(\mathrm{d}y_1) \cdots \pi_K^x(\mathrm{d}y_K), \tag{5}$$

where we wrote the disintegration of π_k with respect to its first marginal μ as $\pi_k(\mathrm{d}x, \mathrm{d}y_k) = \mu(\mathrm{d}x)\pi_k^x(\mathrm{d}y_k)$. With a slight abuse of notation, we will denote $B\#\gamma := B\#\gamma_{1,\cdots,K}$, where

$\gamma_{1,\cdots,K} \in \mathcal{P}(\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_k)$ is the marginal of γ with respect to $(y_1, \cdots, y_K)$. In terms of random variables, if $(X, Y_1, \cdots, Y_K) \sim \gamma$, then $B\#\gamma = \text{Law}[B(Y_1, \cdots, Y_K)]$. Denoting $B\#\Gamma(\mu) := \{B\#\gamma,\ \gamma \in \Gamma(\mu)\}$, we define the multi-valued mapping G which maps $\mu \in \mathcal{P}(\mathcal{X})$ to the set of next iterates $G(\mu) \subset \mathcal{P}(\mathcal{X})$:

$$G := \left\{ \begin{array}{ccc} \mathcal{P}(\mathcal{X}) & \rightrightarrows & \mathcal{P}(\mathcal{X}) \\ \mu & \mapsto & B\#\Gamma(\mu) \end{array} \right. . \tag{6}$$

Taking inspiration from the W_2^2 case, we can see informally the iterate $\overline{\mu} \in G(\mu)$ as a local linearisation of $\mathcal{P}(\mathcal{X})$. To illustrate this intuition, we consider the case $\mathcal{X} = \mathcal{Y}_1 = \cdots = \mathcal{Y}_K$ and assume that for each k, the set of optimal plans $\Pi_{c_k}^*(\mu, \nu_k)$ is reduced to (I, T_k), or in other words, that the Monge problem has a unique solution. Informally, one may see the set of maps $T : \mathcal{X} \longrightarrow \mathcal{X}$ sending μ to a measure $T\#\mu \in \mathcal{P}(\mathcal{X})$ as the tangent space to $\mathcal{P}(\mathcal{X})$ at μ. As a result, the problem of finding a barycentre $\overline{\mu}$ can be seen from the viewpoint of the reference measure μ in the tangent space $T_\mu\mathcal{P}(\mathcal{X})$ as the problem of finding $S \in T_\mu\mathcal{P}(\mathcal{X})$ such that $S\#\mu$ would minimise the cost V. Our approach takes a barycentre of the optimal maps T_k by choosing the candidate $S := B \circ (T_1, \cdots, T_K)$. In the case of the squared-Euclidean cost on the common space $\mathbb{R}^d$, this amounts to $S := \sum_k \lambda_k T_k$, which is exactly the Linearised Optimal Transport barycentre approximation for the reference measure μ, as introduced in [4], Section 4.3. We illustrate this viewpoint schematically in Fig. 1.

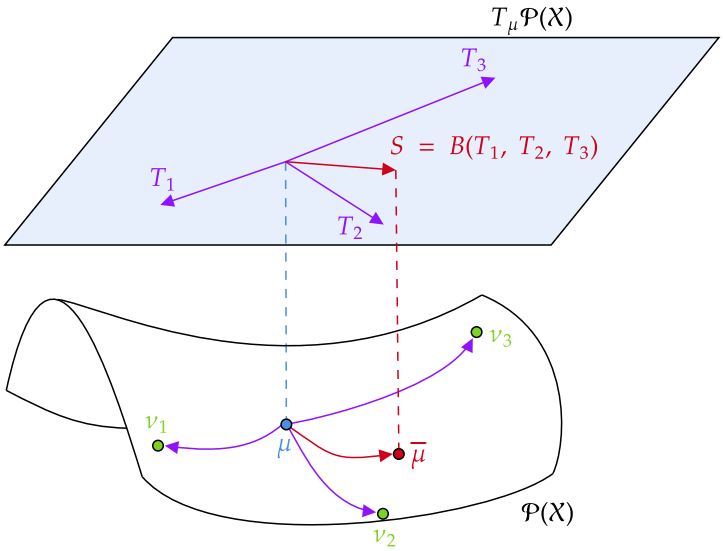


Figure 1: Illustration of the informal linearisation interpretation for the barycentre candidate $\overline{\mu} = B \circ (T_1, \cdots, T_K)\#\mu$.

Starting from a measure $\mu_0 \in \mathcal{P}(\mathcal{X})$, our algorithm consists of choosing iterates through the multi-function G:

$$\forall t \in \mathbb{N},\ \mu_{t+1} \in G(\mu_t).$$

We dedicate the next section to a theoretical study of the convergence of this fixed-point iteration.

Convergence Results

The first main result is that the energy of iterates decreases:

Proposition 1. *Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\overline{\mu} \in G(\mu)$. Then $V(\mu) \geq V(\overline{\mu}) + \mathcal{T}_\delta(\mu, \overline{\mu})$. If μ^* is a barycentre, then $G(\mu^*) = \{\mu^*\}$.*

Along with technical regularity lemmas, Proposition 1 allows us to prove subsequential convergence of the iterates towards a fixed-point of G:

Theorem 2. *For any $\mu_0 \in \mathcal{P}(\mathcal{X})$, let (μ_t) verifying $\mu_{t+1} \in G(\mu_t)$. Then (μ_t) has converging subsequences, all of which converge towards a $\mu \in \mathcal{P}(\mathcal{X})$ such that $\mu \in G(\mu)$.*

These results remain true when modifying the transport costs to their entropic regularisations:

$$\mathcal{T}_{c,\varepsilon}(\mu,\nu) := \min_{\pi \in \Pi(\mu,\nu)} \int_{E\times F} c \mathrm{d}\pi + \varepsilon \,\mathrm{KL}(\pi|\mu\otimes\nu), \quad V_\varepsilon := \sum_{k=1}^{K} \mathcal{T}_{c_k,\varepsilon}(\mu,\nu_k), \tag{7}$$

and taking the multi-plan γ as a gluing of (unique) bi-marginal entropic plans:

$$\Gamma_\varepsilon(\mu) := \mu(\mathrm{d}x)\pi^x_{1,\varepsilon}(\mathrm{d}y_1)\cdots\pi^x_{K,\varepsilon}(\mathrm{d}y_K),\ \forall k \in [\![1,K]\!],\ \pi_{k,\varepsilon} := \Pi^*_{c_k,\varepsilon}(\mu,\nu_k), \tag{8}$$

yielding the iteration (measure-valued) mapping $G_\varepsilon(\mu) := B\#\Gamma_\varepsilon(\mu)$.

A Simpler Variant with Barycentric Projections

For simplicity, we shall consider the case where each $\mathcal{Y}_k = \mathbb{R}^{d_k}$. Drawing inspiration from the simplicity of the case where the couplings are induced by maps, we can define an alternative iteration consisting in choosing a map T_k as the barycentric projection of the coupling $\gamma_{0,k} \in \Pi^*_{c_k}(\mu,\nu_k)$ for $\gamma \in \Gamma(\mu)$: see Definition 3 and Fig. 2.

Definition 3. *The **barycentric projection** of a coupling $\pi \in \Pi(\mu,\nu)$ for $\mu \in \mathcal{P}(E)$ and $\nu \in \mathcal{P}(F)$ is the map $\overline{\pi} : E \longrightarrow F$, which is defined for μ-almost every $x \in E$ as:*

$$\overline{\pi}(x) = \int_F y \pi_x(\mathrm{d}y),$$

where we wrote the disintegration $\pi(\mathrm{d}x,\mathrm{d}y) = \mu(\mathrm{d}x)\pi_x(\mathrm{d}y)$. In terms of random variables, one may write this expression as:

$$\overline{\pi}(x) = \mathop{\mathbb{E}}_{(X,Y)\sim\pi}[Y \mid X = x].$$

We can apply this barycentric projection idea to define an alternate multi-mapping $H : \mathcal{P}(\mathcal{X}) \rightrightarrows \mathcal{P}(\mathcal{X})$:

$$\forall \mu \in \mathcal{P}(\mathcal{X}),\ H(\mu) := \left\{B \circ (\overline{\gamma_{0,1}},\cdots,\overline{\gamma_{0,K}})\#\mu,\ \gamma \in \Gamma(\mu)\right\}. \tag{9}$$

In general, for $\pi \in \Pi(\mu,\nu)$, the push-forward $\overline{\pi}\#\mu$ does not coincide with ν, hence one does not necessarily have $\forall \widetilde{\mu} \in H(\mu),\ V(\widetilde{\mu}) \leq V(\mu)$. However, if each $\Pi^*_{c_k}(\mu,\nu_k)$ are composed of plans supported by maps, then $H(\mu) = G(\mu)$. In the case of discrete measures and for the squared Euclidean cost, the iterations of H correspond to the approach proposed in [2], Algorithm 2.

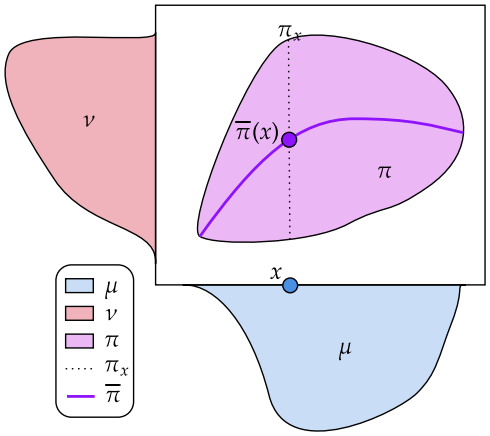


Figure 2: Illustration of a barycentric projection. The disintegration of the coupling π with respect to its first marginal μ at x is the measure π_x concentrated on the dotted line. The barycentric projection of π evaluated at x is the mean of the measure π_x.

Illustration

In this illustration, we look for a barycentre in $\mathbb{R}^2$ whose projections onto different circles match measures on these circles. We choose the costs $c_k(x, y) = \|P_k(x) - y\|_2^2$, where P_k is the projection onto the circle k. Since P_k is not linear, this is a direct generalisation of [3].

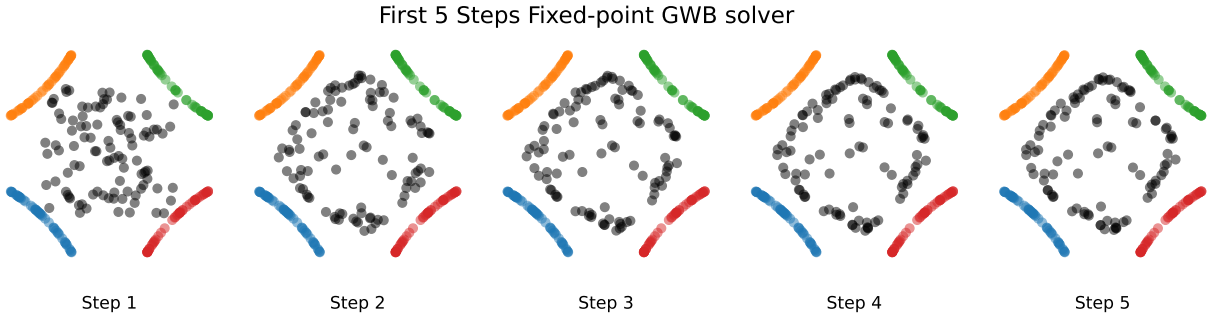


Figure 3: First 5 iterations of the fixed-point algorithm (iterations of H Eq. (9)) for costs $c_k(x, y) = \|P_k(x) - y\|_2^2$, where P_k are projections onto four different circles on which the ν_k are supported (plotted in colour).

In this instance, convergence happens quickly, but a stationary point is only reached after about 5 iterations, as observed on the steps in Fig. 3 and on the energy curve in Fig. 4.

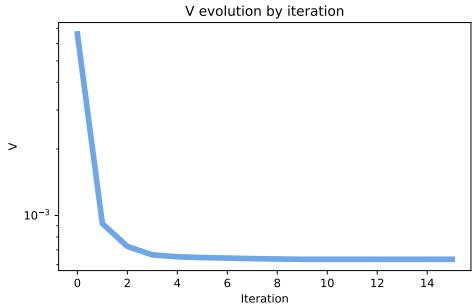


Figure 4: Barycentre energy V of the fixed-point algorithm across iterations.

All code for this experiment and general solvers is available in our companion Python toolkit.

References

[1] Pedro C Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[2] Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Bejing, China, 22–24 Jun 2014. PMLR.

[3] Julie Delon, Nathaël Gozlan, and Alexandre Saint-Dizier. Generalized Wasserstein barycenters between probability measures living on different subspaces, 2021.

[4] Quentin Mérigot, Alex Delalande, and Frederic Chazal. Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space. In *International Conference on Artificial Intelligence and Statistics*, pages 3186–3196. PMLR, 2020.

[5] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[6] Eloi Tanguy, Julie Delon, and Nathaël Gozlan. Computing barycentres of measures for generic transport costs. *arXiv preprint arXiv:2501.04016*, December 2024.

Index des auteurs

Abdesselam Rafik, 1049–1057
Abed Youba, 95–100
Adamczyk-Chauvat Katarzyna, 1181–1186
Adamo Davide, 1578–1582
Adjabi Smail, 1456–1461
Agliz Yasmine, 404–409
Akouba Meliha, 1257–1262
Alahari Karteek, 716, 717
Alberge Julie, 327–332
Alves Paul-Adrien, 585–590
Amara-Ouali Yvenn, 476–483, 1357–1360
Ambroise Christophe, 225–230
Amezouwui Koffi, 1077–1082
Amzal Billy, 1188
Angelim Vieira Felipe Lourenço, 775–781
Aouissaoui Fatma, 1375–1380
Arbel Julyan, 26
Arezki Ouerdia, 1532–1539
Arts Lucie, 652–657
Aubin Jean-Baptiste, 278–281, 1505–1510
Aubinais Eric, 721–726
Aubray Johan, 977–982
Aurouet Daphné, 333–338

Baccou Jean, 484–489
Baillie Nils, 1233–1238
Baitairian Jean-Baptiste, 782–788
Baldé Oumar, 101–106
Baragatti Meili, 1239–1244
Barbieri Antoine, 384–388
Barbot Hugo, 425–430
Barbu Vlad, 498–503
Bassac Francois, 1330–1335
Bastide Paul, 1208–1213
Batton-Hubert Mireille, 283–288
Beaucourt Julien, 1483–1488
Ben Salah Malek, 164–169
Benhamza Manal, 749–754
Benoist Clément, 170–175
Bercu Ariane, 339–344
Berred Alexandre, 627–632
Berthelier Gaspard, 1159–1164
Bertrand Patrice, 1058–1063
Biernacki Christophe, 716, 717
Blanchard Eléonore, 634–640
Bonafos Guillem, 1122–1127
Bonnet Anna, 1138–1145
Bourel Mathias, 1381–1387
Boyer Jérémy, 1399–1405
Bra Jean Armel, 1032–1036
Brault Vincent, 875–880
Brian Britos, 1583–1588
Bruned Vianney, 496
Bry Xavier, 39
Brécheteau Claire, 1406–1411
Brégère Margaux, 46–51, 1159–1164

Cafieri Simona, 272–276, 727–730
Caillebotte Antoine, 990–995
Campagne Eloi, 1009–1013
Cannafarina Hugo, 189–195
Cartier Julie, 431–436
Castelli Luca, 965–970
Casteras Paul, 809–814
Cecchi Nicolas, 1037–1039
Chabassier Agathe, 1128–1137
Chardon Hugo, 207–211
Chaussard Alexandre, 1138–1145
Chauveau Alice, 1336–1340
Chennetier Guillaume, 43, 44
Chevallier Julien, 658–661, 1540–1544
Claeys Emmanuelle, 731–740, 1341–1349
Clerc-Urmès Isabelle, 893–902
Cleynen Alice, 27
Cognot Caroline, 546–552
Cohen Eyal, 212–217
Conanec Nicolas, 1412–1417
Courrier Violaine, 696–701
Cousino Theo, 504–509
Curis Emmanuel, 923–928

De Lara Lucas, 302–307

Debaussart Gwendal, 250–255
Dell'oro Lorenzo, 591–596
Delporte Alexandre, 521–526
Deronzier Mathis, 815–820
Derouet Charlotte, 277
Derquenne Christian, 39, 231–236, 1361–1367
Devijver Emilie, 755–760
Doizé Antoine, 553–557
Doumèche Nathan, 490–495
Drobac Nina, 52–57
Dubois Clara, 1099–1106
Durand Anaëlle, 24, 929
Dusfour-Castan Pauline, 903–913

Ear Philippe, 558–563
Ebrotie Kouassi Alphonse Nathan, 22, 882
Egels Paul, 1462–1469
El Hasnaoui Mohamed, 862–867
El-Boukkouri Fatima-Zahrae, 308–312

Fermanian Jean-Baptiste, 218–223
Fradi Anis, 946–951
Fritsch Coralie, 842–847
Fuentes Vicente Laura, 789–794
Funfschilling Christine, 510–513

Gabaut Auriane, 453–462
Gaetan Carlo, 28, 29
Galharret Jean-Michel, 1214–1219
Gallinari Patrick, 718
Ganassali Luca, 1589–1595
Gannaz Irène, 662–667
Garcia Théo, 527–534
Gerville-Réache Léo, 1368–1373
Gilet Cyprien, 1281–1286
Girard Stéphane, 610–615
Goffard Pierre-Olivier, 641–647
Goffinet Roxane, 58–67
Gomez-Garcia José Gregorio, 616–621
Gomtsyan Marina, 1511–1517
Gorse Nathan, 597–602
Gossard Florian, 1596–1602
Grigoryan Mariam, 1146–1151
Grosdidier Marie, 848–853
Guichard-Sustowski Ketsia, 256–261
Guille Adrien, 719
Guilmineau Camille, 463–468
Guédon Tom, 1189–1196
Guéneau Harold, 1518–1523
Gégout-Petit Anne, 842–847

Haberstich Cécile, 821–827
Hallin Marc, 1040–1042
Helali Salima, 1427–1432
Heranval Antoine, 565–570
Hoayek Anis, 283–288
Hoblos Rouaa, 107–112
Hocking Toby, 30
Huang Yichuan, 1305–1310
Hughes Arthur, 437–445

Jiang Yiye, 1220–1225
Jorge Elise, 446–451
Jouni Jinane, 571–576
Jouvin Nicolas, 828–833

Kaisaridi Sofia, 196–201
Kala Nathanaël, 1165–1170
Kalogeratos Argyris, 250–255
Kammoun Rayhan, 868–874
Kassi Omar, 351–356
Keller Merlin, 1197–1206
Kerioui Marion, 20
Kerleguer Baptiste, 952–957
Kheirallah Rida, 283–288
Khellaf Rémi, 78, 79
Kopp Emma, 1107–1113
Kostic Anica, 1014–1018
Kratz Marie, 883–885
Kuhn Estelle, 31, 32

Lacoste Louis, 668–673
Lacoste Romain Edmond, 1311–1316
Lacour Claire, 33
Lacroix Perrine, 1433–1438
Lafargue Valentin, 731–740
Lamalle Florian, 345–349
Lambert Guerlain, 120
Laplante Félix, 225–230
Lasserre Virginie, 923–928
Laurent Thibault, 1524–1530
Laverny Oskar, 202–205
Le Bars Batiste, 313–318
Le Corff Sylvain, 1138–1145
Le Goff Théotime, 761–766
Le Goff Vincent, 176–181
Lebec Pierre, 1114–1120
Leite Alessandro, 775–781
Lenz Rainer, 741–747
Leterrier Laurie, 577–583
Leveque Nathan, 289–294
Li Zhe, 971–976
Liquet Benoit, 996–1001
Louati Sirine, 1388–1391

Louvet Gaëtan, 1445–1448
Lu Wei, 1287–1293
Luciano Antoine, 1245–1249

Ma Rui, 319–325
Mahmoud Remi, 357–362
Mahoromeza Jean-Luc, 1019–1026
Makhlouf Alexandre, 1545–1549
Mangold Paul, 80
Marrel Amandine, 958–963
Marti Renaud, 854–860
Martineau Louise, 1152–1157
Martinez Herrera Miguel, 1317–1322
Mattei Pierre-Alexandre, 34
Maxime Boucher, 1418–1425
Mayala Moria Grace Aurore, 141–145
Megel Maxime, 237–242
Meneghel Danilevicz Ian, 182–187
Mikem Pierre-André, 81–87
Milhaud Xavier, 395–402
Minchella Paul, 1562–1567
Muller Christophe, 410–417
Métivier David, 546–552

Nassouradine Mahamat Hamdan, 834–840
Ndao Mouhamadou Lamine, 1064–1069
Ndiaye Aminata, 1470–1475
Nedellec Raphaël, 496
Ngatchou-Wandji Joseph, 1375–1380
Niang Seydina Ousmane, 262–269
Nicol Florence, 977–982
Nicolas Maxime, 622–626
Nkameni Daniel, 648–650

Odet Arnaud, 1350–1355
Odonnat Ambroise, 1568–1575
Otogo Ndoutoume Gaetan, 1171–1179
Ouali Maha, 68–76
Oubari Fouad, 702–707

Paillard Joseph, 795–801
Patilea Valentin, 333–338
Peng Liang, 886, 887
Pereira Mike, 1550–1555
Perier Romain, 1392–1397
Petrakos Niki Z., 1263–1271
Pettersson Magnus, 41
Platteau Antoine, 930–937
Principato Guillaume, 1449–1454
Procope-Mamert Sylvain, 1250–1255
Préa Pascal, 1070–1075
Puechmorel Stéphane, 977–982

Quayle Sacha, 939–944

Rabier Charles-Elie, 1226–1231
Razafindrakoto Davidson Lova, 983–988
Razafindralambo Raphaël, 675–680
Reisach Alexander, 767–773
Remiat Justine, 389–394
Reyero Lobo Angel, 418–423, 1002–1007
Rigaill Guillem, 1294–1298
Roblin Arthur, 1489–1494
Rohmer Tom, 469–474
Rolland Antoine, 278–281
Romary Thomas, 1556–1560
Rousseau Judith, 16
Roux Jérôme, 535–538
Runge Vincent, 1299–1303

Sabourin Anne, 35
Sadeler Perrin Maya, 43, 44, 1323–1328
Sadoun Mohamed Djemaa, 1043–1047
Sakho Abdoulaye, 146–154
Sanchez San Benito Alvaro, 1495–1500
Saporta Gilbert, 39
Schall Baptiste, 113–118
Schwartz Antoine, 496
Schürzinger Felix, 1476–1481
Senellart Agathe, 88–93
Serre-Combe Chloé, 603–608
Servien Rémi, 463–468
Sezgin Selman, 363–370
Simon Alice, 243–248
Simon Wood, 36
Sinha Aman, 1576
Smida Zaineb, 371–377
Sokol Harry, 1138–1145
Stocksieker Samuel, 155–162
Strasman Stanislas, 683–688
Stéphanovitch Arthur, 681, 682
Surendran Sobihan, 708–714
Sylvestre Théo, 121–127
Szurewsky Chloé, 1272–1279

Tanguy Eloi, 43, 44, 1603–1608
Taquet Vianney, 496
Temple-Boyer Marie, 514–519
Thulliez Emma, 539–544
Tio Maxime, 23, 914
Toti Vito, 23, 914
Tous Jeanne, 270
Trufin Julien, 37

Valla Mathias, 395–402
Van Bever Germain, 378–382

Van Biesbroeck Antoine, 128–133
Van De Geer Sara, 18
Vaucher Rémi, 295–300
Vayness Eyal, 1439–1443
Verdier Claire, 1027–1030
Vergne Nicolas, 1083–1091
Vernet Elodie, 875–880
Verronneau Rémi, 1092–1097
Vidagbandji Lucien M., 627–632
Voinot Charlotte, 802–807
Walker Emily, 848–853
Weinberger Simon, 134–139
Wintenberger Olivier, 888–891
Wu Yu-Han, 689–694

Yao Anne-Françoise, 977–982
Yapi Aho, 915–921

Zehavi Itai, 1501–1503

Centrale Méditerranée
INSTITUT de MATHÉMATIQUES de MARSEILLE
BUREAU DES CONGRÈS AIX-MARSEILLE PROVENCE
VILLE DE MARSEILLE
Digital insurance and long term risk
Chaire d'Excellence
CNP
Chaire ACTIONS
CARDIF
GROUPE BNP PARIBAS
PELLENC ST. aMias
LES MATHS VECTEUR D'INNOVATION
FRUMAM
Fédération de Recherche des Unités de Mathématiques de Marseille –
amU
Aix Marseille Université
56èmes Journées de Statistique
de la SFdS à Marseille
du 2 au 6 juin 2025