MÉMOIRE

Pour obtenir le diplôme

HABILITATION À DIRIGER DES RECHERCHES
DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et Technologies de l'Information, Informatique
Spécialité : Mathématique Appliquée
Unité de recherche : Laboratoire d'Informatique de Grenoble

Contributions en apprentissage statistique pour des données structurées en grand dimension

Contributions to Statistical Learning for Structured Data in High-dimension

Présentée par :

Emilie Devijver

Rapporteur·rices :

Florence d'ALCHÉ-BUC
Professeure des universités, Télécom Paris

Antoine CHAMBAZ
Professeur des universités, Université Paris Cité

Arthur CHARPENTIER
Professeur des universités, Université du Québec à Montréal

Habilitation soutenue publiquement le **10/12/2024**, devant le jury composé de :

Sophie ACHARD Directrice de recherche, CNRS	Examinatrice
Florence d'ALCHÉ-BUC Professeure des universités, Télécom Paris	Rapportrice
Antoine CHAMBAZ Professeur des universités, Université Paris Cité	Rapporteur
Julie JOSSE Directrice de recherche, INRIA	Examinatrice
Pascal MASSART Professeur des universités, Université Paris Saclay	Examinateur
Jean-Michel POGGI Professeur des universités, Université Paris Saclay	Examinateur

Remerciements

Ce manuscrit rassemble une partie de mes travaux de recherche réalisés depuis ma thèse, entre 2015 et 2024. Il est aussi le fruit de 9 années passées dans l'enseignement supérieur et la recherche (hors thèse). J'ai profité de cette opportunité pour faire un point introspectif, et voici quelques mots résumant mon point de vue. Bien que ce milieu puisse être solitaire, exigeant et parfois sournois, j'ai eu le bonheur d'être entourée de personnes bienveillantes. Ces personnes m'ont aidée à définir mes limites et mes objectifs, tout en me permettant de faire de la science dans un environnement motivant et chaleureux. Je me permets de prendre un peu de place dans ce manuscrit pour remercier ces collègues, étudiants et amis (intersection non vide), sans qui ma vie professionnelle serait bien moins passionnante !

Pour commencer, je tiens à remercier mes rapporteurs pour leur temps et leur mots encourageants. Antoine, Arthur et Florence, vous êtes pour moi des modèles de gentillesse, de curiosité scientifique et de grande rigueur : tous les ingrédients (ordonnés) du chercheur parfait. Je suis profondément honorée et touchée que vous ayez pris le temps de rapporter ce manuscrit, et qu'Antoine et Florence soyez présents pour la soutenance. Un remerciement particulier à Arthur, qui a accepté cette lourde tâche dans un délai très court. À un moment où je traversais une série de déconvenues, tu as incarné l'idée que le monde de la recherche peut rimer avec fraternité. Merci Antoine, pour ton enthousiaste, pour les JES en 2018 et pour ta générosité. Florence, depuis 2017, nos rencontres, bien que sporadiques, ont toujours été une source de richesse intellectuelle et d'inspiration. Une simple discussion avec toi occupe et anime mon esprit pendant des semaines. Merci.

Je tiens également à remercier les membres du jury pour leur participation à la soutenance. Jean-Michel et Pascal, en tant que directeurs de thèse vous avez été mes premiers guides , et je suis très touchée (et un peu stressée, je dois l'avouer) d'avoir votre regard aujourd'hui sur mes travaux. J'entends très régulièrement votre petite voix pour me guider quand je suis un peu perdue (dans un calcul ou face à une situation inhabituelle). Merci.
Julie, merci pour ton temps, et les échanges ces dernières années. Sophie, merci également pour les discussions enrichissantes et enthousiasmantes que nous avons eues au fil des années.

Ce manuscrit rapporte une partie de mes travaux, certains n'ayant pas abouti à un formalisme suffisant pour y figurer, d'autres étant thématiquement trop éloignés. Néanmoins, je tiens à remercier sincèrement toutes les personnes avec qui j'ai discuté de science, de près ou de loin. Ces remerciements sont avant tout professionnels, même si bon nombre des personnes mentionnées ici sont bien plus que de simples collègues.
Merci Mélina, pour notre équipe de `shock` et pour tout le reste, les coups de pression et les coups de passion. Cheerleader un jour, cheerleader toujours.
Merci Emeline, pour ton soutien, nos cafés entre deux trains, et les regressions inverses. J'ai hâte de voir ce que l'avenir nous réserve !
Merci Irène et Gerda de m'avoir accueillie en Belgique. J'ai adoré la vie flamande, mais surtout votre rigueur scientifique et votre curiosité insatiable. Ma vie de chercheuse aujourd'hui est fortement influencée par ces deux années à Leuven.
Merci Valérie, pour cette rencontre fortuite lors d'un événement social où ni toi ni moi ne souhaitions aller. Ce hasard a mené à de nombreuses discussions et collaborations autour de modèles mixtes mal posés et de données réelles complexes. Merci également à Madison, dont j'admire la persévérance et la rigueur, et à Marie pour notre travail ensemble sur les données NASH.

Merci Adeline, pour ta générosité depuis notre première discussion sur les EM aux JdS alors que j'étais encore en thèse. Tu m'as montré que le monde de la recherche n'est pas cloisonné entre doctorants et permanents. Tu as cru en moi, m'a aidée lors des candidatures, et m'as accompagnée dans mon intégration et mon évolution depuis 2017, aussi bien professionnelle que personnelle. Tu es la personne à qui je peux demander de l'aide sur tout, que ce soit une preuve, un code, un enseignement, un appel à projet, ou encore cette HDR (et j'en ai eu besoin !). Nos discussions m'ont aussi énormément fait progresser sur des sujets qui me sont chers, comme les inégalités de genre dans la recherche, l'équilibre vie pro/perso, et les choix de carrière. J'ai beaucoup de chance de t'avoir près de moi.
Merci Vincent, pour toutes tes casquettes, au sens propre comme au sens figuré (mon placard te remercie moins). Modèles de mélange, programmation, coaching de vie pro, psy... la liste est longue. Tu m'es un soutien précieux.
Merci Marianne, pour ce conseil au FEM 2016 de contacter le LIG, puis pour m'avoir poussée à faire de l'analyse fonctionnelle sur les PR trees. Merci de chercher avec moi qui cause quoi, on s'amuse bien. Tu es la meilleure organisatrice de goûter, mais c'est surtout pour ton énergie et ta grande curiosité que je multiplie nos collaborations !
Merci Massih, pour ton accueil incroyable dans l'équipe, et pour les innombrables discussions scientifiques. Pour les cafés bien sûr, mais surtout pour ton regard toujours bienveillant, et de m'avoir épaulé pour mon premier co-encadrement. Tu as toujours tout fait pour créer les meilleures conditions, à la fois pour moi et pour l'équipe. Pas facile de reprendre le flambeau ! Merci de m'avoir invitée à de nombreux projets, je n'ai jamais manqué de nourriture, scientifique ou alimentaire.
Merci Eric, pour l'accueil et les 1001 projets. Pour le premier co-encadrement aussi, et pour m'avoir entrainé à travailler en causalité. Pour ta pédagogie sans faille, même lorsqu'il s'agissait de me réexpliquer pour la centième fois les architectures de deep learning ou les subtilités des modèles de traitement automatique de la langue. J'apprécie ton habileté à te mettre au niveau des autres, et à tirer le meilleur de chacun de nous.
Merci Ahlame, Aude, Christophe, Gilles, Patrick, pour votre gentillesse et bienveillance au quotidien. Vos sourires et nos échanges font du labo un endroit où j'arrive toujours avec plaisir le matin. Merci Cécile pour les discussions sans fin pour changer le monde.
Merci Charlotte, pour ton accueil chaleureux à Grenoble, le cooking club, mais surtout merci pour tous nos échanges depuis, scientifiques ou non, mais toujours inspirants.
Merci à Charles et Vasilii, mes premiers doctorants (quasi jumeaux), pour votre confiance. Avoir co-encadré votre thèse a été une expérience enrichissante et profondément joyeuse, et c'est aujourd'hui mon activité favorite : travailler avec des doctorants. Vasilii, nos différences d'organisation n'ont jamais entravé une collaboration réussie. Charles, ta ténacité – tu es parfois têtu, il faut l'admettre ! – fait avancer la science. Vous m'avez tous deux montré que l'encadrement doctoral était une source infinie de partage et d'apprentissage, et j'espère vous avoir formé à la recherche, notamment en terme de rigueur et de formalisme. En tout cas, je suis fière de vos parcours. Merci Lei, pour avoir pris la relève et trouvé ta place dans notre trio parfois envahissant. J'ai hâte de voir où l'avenir te mènera.
Nous avons eu la chance d'avoir l'institut 3IA MIAI, et j'ai pu en bénéficier au travers de deux chaires : merci Massih et Alexis, et Adeline et Anatoli.
Merci Noel et Roberta, pour la collaboration interdisciplinaire. Il nous fallu du temps pour nous comprendre, mais j'ai découvert des problématiques statistiques très intéressantes grâce à nos projets. Merci Sébastien d'avoir été le premier thésard à faire le pont. Thank you, Ashna and Johannes, for trusting me as your ML supervisor. Ashna, I am very proud of your learning curve in statistics. Johannes, I am truly impressed by your technical expertise. Joao, merci pour ta joie de vivre, et tes capacités de modélisation et d'implémentation.
Merci à Marta, pour notre projet sur les bandits et pour les sorties avec les enfants.
Encadrer des postdocs est une aventure à part, pleine de défis. Merci à Myriam, Anouar, Lila et Daria pour votre confiance. Trouver la juste place entre vous accompagner dans votre épanouissement scientifique et mettre les mains dans le cambouis n'a pas toujours été simple. Daria, merci tout particulièrement de m'avoir ouvert les yeux sur des questions fondamentales mais aussi sur les subtilités liées aux données.
J'en profite pour remercier Gregor pour les différentes collaborations stimulantes, où nos perspectives complémentaires m'ont bien fait progresser scientifiquement. Merci Maximin, pour

nos discussions sur les modèles de langue. Merci Maxime, pour le point de vue complémentaire sur l'abstraction de graphes.
Merci à Sami pour ton incroyable travail d'implémentation, ce fut un réel plaisir de travailler avec toi. Merci aussi à Alexandre et Lorys pour votre engagement sur le projet.
Merci Georges pour toutes les discussions scientifiques, qui me font cogiter pendant des semaines.
Merci à Jean-Charles et Annique pour notre collaboration sur les données de *mouse-tracking*, j'ai beaucoup appris des problèmes de modélisation. Merci à Pascal, Julie, Christophe et Sébastien pour notre travail sur les données de cancérologie et d'apnée du sommeil. Comme pour la science des matériaux, ces applications m'ont poussée à explorer des modèles et même une bibliographie que je ne connaissais pas.
Au présent, un sincère merci à Théotime, Alexander, Vsevolod, Clément, Flora, Morad pour votre confiance et pour ce bout de chemin à parcourir ensemble. J'espère vous apporter ce dont vous aurez besoin. Merci également à tous les stagiaires et doctorants de l'équipe que je n'ai pas mentionnés, mais à qui je pense en rédigeant ces lignes. La vie au sein de l'équipe est belle grâce à vous tous.
Thank you to CoCaLa, the causality team in Copenhagen, for their warm welcome during spring 2024. This manuscript was finally completed while I was there. I learned a lot on the statistical side, both from Jeff's thesis and the seminars. I thoroughly enjoyed the retreat, where I discovered many interesting papers, and which served as a model for team building while doing science.

Merci à toute l'équipe ADMINFI d'être en soutient de notre travail, et particulièrement Latifa et Samantha pour l'aspect équipe, Michèle pour l'aspect RH, et Stéphanie pour les budgets. Je ne sais pas comment c'était avant, mais j'ai vu l'évolution ses dernières années, et vous faites un travail formidable malgré les conditions qui se dégradent.
Un grand merci à Amiel, de l'école doctorale, pour m'avoir accompagnée pour cette habilitation.

Pour conclure ces remerciements, un mot pour ma famille, qui supporte, bon gré mal gré, ma vie professionnelle un peu envahissante. Merci aux Orsay girls, d'être là au jour le jour depuis tant d'années. Merci à Elodie et Pierre, pour les diners et les randos. Merci à Benoit et Fabien, pour les discussions plus ou moins scientifiques. Merci, maman et papa, d'essayer de comprendre ce que je fais, et merci pour votre soutien infaillible. Merci Pascale et Jean-Claude pour votre curiosité. Merci Rémi, pour tout, pour toi. Et merci, Paul et Eva : pas pour les nuits trop courtes, mais pour tout le reste. Je vous aime.

Introduction

Embarking on an academic career is like beginning a journey through uncharted territory, where each step forward brings new discoveries, challenges, and opportunities for growth. As I reflect on the trajectory of my research since completing my Ph.D., I am pleased to see the many experiences that have shaped my path over the past nine years. In this manuscript, I aim to put together some important threads of this journey.

While working on this introduction, I have wondered how best to present the diverse array of topics I have worked on. Ultimately, I realized that the most authentic narrative emerges from the relationships forged and the collaborations developed along the way. It is through the lens of these connections—with colleagues, many of whom have become friends—that the true essence of my journey as a researcher comes into focus.

Through this narrative, I hope to offer not only an overview into my personal and professional evolution but also a testament to the significant power of community and collaboration.

Once upon a time ...

My research journey in statistical learning began with my PhD thesis, defended in July 2015, which was supervised by Pascal Massart and Jean-Michel Poggi. Entitled *High-dimensional mixture regression models, application to functional data*, my thesis dealt with the challenges of model selection, specifically the slope heuristics (Birgé and Massart, 2001; Arlot, 2019), for Lasso penalty in **mixture regression models in high-dimension** (Devijver, 2015a,b), alongside its low-rank counterpart for multivariate output (Devijver, 2017a). We proposed a method for model-based clustering, extended to **functional data** (Devijver, 2017b), later applied in electrical datasets (Devijver et al., 2020).

Following the manuscript's submission to reviewers, I started to work with Mélina Gallopin, a fellow PhD student in my lab at the time, marking the beginning of a fruitful partnership and enduring friendship. This collaboration focused on the slope heuristic for **Gaussian Graphical Models**, specifically on detecting block-diagonal patterns using thresholding techniques applied to empirical covariance matrices (Devijver and Gallopin, 2018). Mélina was working on this topic in her PhD thesis, and we discussed it through the Select seminar. I brought my knowledge in theoretical tools for model selection. Presenting this work at a seminar, Christophe Biernacki asked, among others questions, to discuss the **stability** of the method. In practice, it was clear that the method was stable by construction, without resampling, but in theory, it was less clear. Recently, with Mélina Gallopin and Rémi Molinier, we proved the stability of the method using topological tools (Devijver et al., 2024).

During my postdoc at KU Leuven, I further explored the field of functional data analysis, working alongside Gerda Claeskens and Irène Gijbels on modeling the **misalignment of functional data** (Claeskens et al., 2021). This project led to the development of a nonlinear mixed-effect modeling approach to account for phase and amplitude variability, yielding a consistent estimator- a formidable challenge in both modeling and theoretical analysis. I learned about mixed models, asymptotic theory, and went deeper into functional data modeling. It was also my first international experience, giving me a new insight into research and teaching.

During my postdoctoral tenure, I was looking for an academic research position, a process that demanded considerable time and patience. While these two years were undeniably

challenging, I took advantage of them by actively engaging with other talented researchers to explore new and diverse directions.
In discussions with Emeline Perthame about her postdoctoral research, we started to work on the underlying **model of inverse regression**. Together, we proposed prediction regions (Devijver and Perthame, 2020) aimed at elucidating the confidence in estimates compared to classical methods in high dimensions. It was, for Emeline and me, our first project *alone*, without any mentor. It relied on Emeline's knowledge of the model, and my theoretical knowledge of non-asymptotic and asymptotic tools.
At the same time, Emeline and I worked with Mélina Gallopin to extend Emeline's method of mixture of inverse regression using our graphical models to consider nondiagonal covariance matrices. Our focus centered on addressing biological challenges, with the development of a novel method that combines the strengths of both models to predict quantitative trait outcomes from biological data (Blein-Nicolas et al., 2024), in collaboration with Mélisande Blein-Nicolas.
At that time, I also had the opportunity to meet Valérie Monbet, an encounter that significantly influenced my research direction. I joined a project with Valérie, Madison Joyce Giacofci and Marie Morvan about *mixture of logistic regression models for functional data*. Although the modeling aspect was reminiscent of my PhD work, the approach was completely different, driven by near-infrared spectrometry data aimed at predicting Nonalcoholic Steatohepatitis (NASH) (Morvan et al., 2021).

Then, I get a researcher position at CNRS in mathematics, working in a computer science laboratory. In October 2017, I started my tenure at the Laboratoire d'Informatique de Grenoble[1], where I found myself amidst a plethora of diverse ongoing projects, and there will soon be many more in the pipeline. Indeed, upon joining the Aptikal team (formerly AMA team), I was warmly welcomed by my new colleagues, both socially and scientifically.

Massih-Reza Amini and I started several exciting projects. The first one, on **semi-supervised learning in high-dimension**, was my first research project supervising a PhD student, with the support of Massih: writing the project proposal, securing funding, meeting the Master 2 students, supervising an intern and proposing the PhD. Vasilii Feofanov was the student we selected, which blossomed into a fruitful collaboration (Feofanov et al., 2022, 2024).
Massih also involved me in interdisciplinary projects, among which the MAGNET chair of the MIAI institute, which led to collaborations with Noel Jakse and Roberta Poloni from SIMAP on machine learning for material science. We started, with Noel and Rémi Molinier, with the PhD thesis of Sébastien Becker, where we used existing ML tools and topological data analysis to understand the crystallization of monoatomic metals (Becker et al., 2022). Our collaboration then evolved to include the development of new ML models and methods for material science challenges: with Noel through the PhD thesis of Johannes Sandberg, where we use **feature selection penalization within the high-dimensional neural network potential** to detect relevant descriptors (Sandberg et al., 2024), and with Noel and Roberta through the PhD thesis of Ashna Jose, where we develop an **active learning** method to construct the training set for MOFs discovery (Jose et al., 2023, 2024). In these three projects, I was the only statistician in the supervising team.

In the meantime, Eric Gaussier invited me to participate in two projects. The first involved **probabilistic regression trees**, in collaboration with Marianne Clausel and working with Myriam Tami as a postdoc, Sami Alkhoury and Alexandre Seiller as engineers (Alkhoury et al., 2020). It can be seen as an extension to standard regression trees[2], but needs tools from functional analysis for the theoretical side.
The second project was the CIFRE PhD project of Charles Assaad on causal discovery for time series (Assaad et al., 2022a,c). Since then, **causality** has become one of my primary research interests. While my prior knowledge in graphical models served as a foundation, my collaboration with Eric (and later Gregor Gössler) has allowed me to explore causality further, par-

[1] It is worth noting that the applied mathematics lab, Laboratoire Jean Kuntzman, shared the same building. I found it anecdotal when preparing the concours, it was in fact fundamental to keep the link with my colleagues in statistics, and facilitate my work to construct a bridge between the two labs.
[2] which I was familiar with, thanks to Jean-Michel Poggi.

ticularly in the context of time series analysis: we worked on causal discovery for mixed time series through Lei Zan's PhD (Zan et al., 2022), and on **causal reasoning for time series** within Anouar Meynaoui's postdoc (Assaad et al., 2024). Recently, I received a junior chair at the MIAI institute, where I aim to bridge classical statistics and causality, with application to healthcare (more details on this will be discussed in the conclusion/perspectives section).

Since I arrived in Grenoble, I also collaborated with researchers from Laboratoire Jean Kuntzman on several projects. With Vincent Brault and Charlotte Laclau, I explored **mixture of segmentation models** (Brault et al., 2024), which brought me back to the realm of mixture models. We particularly devoted this method to functional data, where segmentation makes a lot of sense. Additionally, I had the opportunity to collaborate with Adeline Samson on **simultaneous confidence bands** (Devijver and Samson, 2024). Our collaboration began with theoretical questions during my postdoc and evolved into a detailed exploration of modeling approaches, particularly the assumptions on bias that are usually made in the literature but are rather strong.

Manuscript outline

This journey may seem chaotic, but there are several ways to cluster the projects.
From the modeling perspective, mixture models, functional data/time series, and networks have been the primary objects of my study; I like to think of it as an **exploration of structured data, mainly in high-dimension**. I am also interested in several aspects of **uncertainty**, through the statistical inference of simultaneous confidence bands or prediction intervals, modeling the uncertainty in the covariates by introducing probabilistic trees, or considering abstract graph in causal inference.
In terms of mathematical tools, I have engaged with concentration inequalities, functional analysis, penalized estimators, graph theory, topological data analysis and asymptotic statistics. Even though these topics may seem unrelated, I have enjoyed exploring the connections between them, with progress in one project often providing insights into another—sometimes even beyond what was initially planned.
I would also like to mention that I have had the privilege of collaborating with experts from various fields, including biology, material science, and healthcare. These collaborations may not always lead to statistical projects, but they have certainly enriched my understanding and perspective. My goal has always been to propose interpretable models, which facilitate interdisciplinary collaborations.

This manuscript is organized into four fundamental chapters, and an applied chapter:

Chapter 1 **Methods for High-Dimension Regression**. This chapter addresses the regression task, where we explain a continuous response from covariates. This is a vast field of research in statistics and machine learning, but in this manuscript, we particularly focus on: 1) the classical univariate/multivariate response, with contributions in modeling a nonparametric prediction function with uncertainty in covariates and modeling the uncertainty around the prediction for a fully parametric model in high dimensions; and 2) functional data modeling, where we provide several new models validated by theoretical results.

Chapter 2 **Network Inference by Gaussian Graphical Model and Its Use**. This chapter focuses on the dependence between covariates. We first theoretically address the estimation of the Gaussian graphical model through a block-diagonal structure and discuss the stability of this estimator. Then, we propose some prediction models based on this network structure, driven by real problems in biology, where the contributions are mainly practical.

Chapter 3 **Causal Inference for Time Series**. This chapter begins with an introduction to causal inference, particularly for time series, which is my main research focus. We then introduce our contributions: 1) several independence measures for different kinds of data; 2) causal discovery methods with various experiments; and 3) causal reasoning through the identifiability on an abstract causal graph. The contributions here are methodological and theoretical.

Chapter 4 **Semi-Supervised Learning**. This chapter deals with the semi-supervised paradigm, where many unlabeled points and few labeled points are observed. Two specific tasks are considered: 1) constructing the labeled set from scratch in both regression and classification contexts, known as the active learning problem; and 2) multi-class classification with partially labeled data, where we propose an algorithm to construct pseudo-labels for unlabeled points to improve classification performance. The contributions here are methodological and theoretical.

Chapter 5 **Application in Material Science**. This chapter focused on applications of statistical learning in Material Science to highlight some significant practical contributions.

As each chapter is independent of the others, I have provided an introduction for each, so I will not give more details in this section. Each chapter begins with a summary of my contributions, depicted as a list of collaborators and papers. Then, a broad introduction is provided, motivating the context and discussing the state-of-the-art. Finally, selected contributions are detailed. Details are omitted, as the goal of this manuscript is to provide a broader view of my research.

Finally, the manuscript concludes with a discussion on various perspectives and directions for future research. The main point is that this journey is not finished, and I am excited about what is next!

Contents

Chapter 1

Methods for high-dimension regression

This chapter is the result of collaborations with Gerda Claeskens and Irène Gijbels (KU Leuven), Emeline Perthame (Institut Pasteur), Marianne Clausel (IECL, Mathematics Research Institute), Eric Gaussier (LIG, Computer Science Laboratory), Vincent Brault (LJK, Applied Mathematics laboratory), Charlotte Laclau (Télécom Paris), Adeline Samson (LJK, Applied Mathematics laboratory) and Myriam Tami (Postdoc student), Sami Alkhoury (Engineer) and Alexandre Seiller (Engineer). Thanks to them!

- Should we correct the bias in Confidence Bands for Repeated Functional Data?, Devijver, E. and Leclercq, A. (2024), preprint, link HAL.
- Ensembles of PR trees, Alkhoury, S., Clausel, M. Devijver, E., Gaussier, E. and Seiller, A. (2024), preprint, link HAL.
- Mixture of segmentation for heterogeneous functional data, Brault,V., Devijver, E. and Laclau, C. (2024), Electron. J. Statist. 18(2): 3729-3773 link.[a]
- Nonlinear mixed effects modeling and warping for functional data using B-splines, Claeskens, G., Devijver, E., and Gijbels, I. (2021), Electronic Journal of Statistics, 15(2): 5245-5282, link.[b]
- Smooth and consistent probabilistic regression tree, Alkhoury, S., Devijver, E., Clausel, M., Tami, M., Gaussier, E., and Oppenheim, G. (2020). In Advances in Neural Information Processing Systems 34, link.[c]
- Prediction regions through inverse regression, Devijver, E. and Perthame, E. (2020). Journal of Machine Learning Research, 21(113):1–24, link.

[a] Code available at https://github.com/laclauc/MixtSegmentation
[b] Code available at https://cran.r-project.org/web/packages/warpMix/index.html
[c] Code available at https://gitlab.com/sami.kh/pr-tree

Regression task corresponds to describe a (potentially mutlivariate) continuous response $\mathbf{Y} \in \mathcal{Y}$ by some covariates $\mathbf{X} \in \mathcal{X}^D$. This has been studied through many models, from the basic linear heteroscedastic to the complex nonparametric one.

In this chapter, we particularly differentiate two cases: when $\mathbf{Y}$ is an univariate or a multivariate variable, and when $\mathbf{Y}$ corresponds to a function. Methods behind are different: in the first case, we use different covariates to explain the response, while in the second case, we use a functional basis to describe the function. If the model seems to coincide, and the standard estimators can be used in both cases, the differences are of our interest.

The organization of this chapter is the following:

- In Section 1.1, we focus on the regression task with univariate or multivariate variables. We propose two contributions. The first, detailed in Section 1.1.1, focuses on the linear regression model in high dimensions. We propose to evaluate the uncertainty around the prediction by constructing prediction regions when using the inverse regression trick for the estimation. This is a joint work with Emeline Perthame, and all the details are available in Devijver and Perthame (2020). The second, detailed in Section 1.1.2, is a nonparametric estimator based on the standard regression tree, called probabilistic regression trees, which consider noise in covariates to smooth the prediction function. Ensemble methods, like boosting, bagging and bayesian additive models, are discussed, and consistency of all these estimators is provided. This project is a joint work with Marianne Clausel and Eric Gaussier. It starts with the PostDoc of Myriam Tami, who works on an initial version; and Sami Alkhoury and Alexandre Seiller worked on the numerical experiments. Details for one probabilistic regression tree are available in Alkhoury et al. (2020), while the ensemble versions are discussed in Alkhoury et al. (2024).

- In Section 1.2, we focus on functional data. When projecting the functional data onto a functional basis, the statistical analysis returns to a regression task over the functional basis. We propose three contributions in this manuscript. First, we propose a model in Section 1.2.1 that takes into account the variability in phase and in amplitude through a nonlinear mixed effect model. This was the topic of my postdoctoral study, in collaboration with Gerda Claeskens and Irène Gijbels. All the details are available in Claeskens et al. (2021). Then, in Section 1.2.2, we address another modeling problem concerning heterogenous data modeled into homogeneous clusters and homogeneous regimes. This is a joint work with Vincent Brault and Charlotte Laclau, and more details are available in Brault et al. (2024). In both models, we provide a theoretical analysis about the identifiability of the model and the consistency of the proposed estimators. Finally, in Section 1.2.3, we study the uncertainty on the estimation of the mean curve by providing simultaneous confidence band for the linear model. We particularly recast this as a model selection problem. This is a joint work with Adeline Samson, and more details are available in Devijver and Samson (2024).

1.1 Regression with univariate and multivariate response

Many methods and models have been proposed for the regression task in the literature. This section introduces two contributions with different modeling but answering similar questions. Let consider some covariates $\mathbf{X} \in \mathcal{X}^D$, with D large, and $\mathbf{Y} \in \mathbb{R}^L$ the output, and a sample $(\mathbf{x}_i, \mathbf{y}_i)_{1 \leq i \leq n}$ from those random variables of size n.

When considering **linear models**, if n is large, with respect to D and L, the standard least square estimator has good performance, but it can be problematic for interpretation. Regularized regression reduces the dimension of the regression problem to the subset of the most relevant features. Methods include the Lasso (Tibshirani, 1996), the Dantzig selector (Candes and Tao, 2007), or the Ridge estimator (Hoerl and Kennard, 1970) to refer to the most popular. These widely used methods are designed to account for univariate response and few implementations exist for multivariate response, considering then independent response terms. Some extensions have been proposed for generalized linear models, as introduced for example in Buhlmann and van de Geer (2011). Another way to deal with high dimensional data consists in dimension reduction techniques which extract components or latent variables that summarize the information of a large dataset into a small dimension space. For example, the Principal Component Regression (PCR) selects a subset of principal components for regression and focuses on hyperplanes; the Partial Least Square regression (PLS) projects the predicted variables and looks for latent variables, correlated to both response and covariates, in order to perform the regression of $\mathbf{Y}$ on $\mathbf{X}$ in a space of lower dimension than D ; and the Sliced Inverse Regression (SIR) introduced in Li (1991) restricts the regressors to few projections by inverting the role of predictors and response. SIR is based on a prior linear dimension reduction by considering the covariance matrix of the inverse expectation $\mathbb{E}(\mathbf{X}|\mathbf{Y})$ (hence the name of the method). The main assumption of SIR relies on Linearity Design Condition, satisfied by elliptical distributions. However, the number of axes to retain must be specified beforehand, which is one of the main drawbacks of those methods. More precisely, in the context of regression with random predictors, several authors proposed reduction dimension techniques based on the joint distribution of both predictors and response (George and Oman, 1996; Helland, 1992; Helland and Almøy, 1994) to identify components used to reduce the dimension of predictors matrix. Interestingly, while the regression of interest usually models the conditional distribution of response given predictors $\mathbf{Y}|\mathbf{X}$, some authors explored the properties of inverse models, meaning that the conditional distribution of predictors is studied given the response $\mathbf{X}|\mathbf{Y}$ (Oman, 1991). See Cook (2007) for an interesting overview of these techniques. Whereas variable selection methods are mainly used for high-dimensional data, the inverse regression approach is particularly interesting in three specific frameworks. First, when $D >> N$, if a large number of covariates is known to have an impact on the response, selecting variables is not relevant while inverse regression is effective. Secondly, when dealing with large dimension for both sample size and number of predictors (N and D large), inverse regression is efficient under few weak assumptions: it avoids the inversion of a large empirical covariance matrix which is time consuming in practice even if it is invertible in theory. Thirdly, inverse regression has the advantage to allow multiple response potentially correlated, which is more and more frequent with real data.

When considering **nonlinear models**, either the link function can be explicit, using polynomial functions, or any precise combination of functions, or nonparametric models are considered, to not define explicitly the model (Hastie et al., 2001). Several nonparametric models have been proposed, such as kernels and neural networks, but regression trees (Breiman et al., 1984) and the ensemble methods based on them such as random forests (Breiman, 2001), gradient-boosted trees (Friedman, 2000; Elith et al., 2008) and Bayesian additive regression trees (Tan and Roy, 2019) have been successfully used for regression problems in many applications. For regression trees, the feature space is partitioned into a set of hyper-rectangles, and a constant model is fitted in each region. As a result, standard regression trees may have difficulty adapting to the smoothness of the link functions and the noise in the input data. Extensions of regression trees have been proposed to generalize this prediction function. Soft trees (Irsoy et al., 2012) and fuzzy trees (Suarez and Lutsko, 2003) are both used for classification and regression

and can learn a parameter vector at each node, the dimensionality of which is equal to that of the input data. For a specific node, this vector is used in a gating function: it gives the probability for each observation to be assigned to the left children of the node. Each example is thus assigned to all leaves with a certain class membership, and the final prediction is a smooth combination of the prediction at each node. Soft and fuzzy trees can be seen as a direct extension of the hierarchical mixtures of experts (HME, Jordan and Jacobs (1994)): indeed, if the HME use predefined trees or trees learned from another method (typically, a standard decision regression tree), then soft trees are constructed based on the hierarchy of experts. Smooth transition regression (STR) trees, introduced in da Rosa et al. (2008), follow the same general principle but instead rely on a single parameter at each node. A sigmoid-based gating function is also used to assign points to different regions of the tree. Instead of focusing on one tree, ensemble methods have been proposed to improve regression and classification tasks. The most well-known ensemble methods based on regression trees are certainly random forests (RF) introduced in Breiman (2001), where small trees are averaged to reduce the variance, and gradient-boosted trees (GBT) (Friedman, 2000; Elith et al., 2008) with an additive method, where each new tree reduces the resulting error, thus reducing the bias. Soft trees, STR trees, and PR trees, viewed as construction blocks, can also be used in ensemble extensions to reduce the bias or variance and thus improve the global performance. More recently, the ensemble method of Bayesian additive regression trees known as BART (Chipman et al., 2010) has been proposed. As an ensemble method, many trees are combined. The Bayesian *a priori* is used to define the structure of each tree and the parameters necessary to define each one. The boosting model is used to reduce the error, although an overall average is also calculated to reduce the variance. The extension to soft trees has been proposed, namely soft-BART (Linero and Yang, 2018), which also allows for sparsity using a Dirichlet *a priori* on the feature space.

When considering prediction, one can be interested directly in the value of the prediction, or to measure the uncertainty around this estimator. From a theoretical viewpoint, consistency of the prediction has been achieved for many models: regression trees (Györfi et al., 2002), standard RF (Scornet et al., 2015), boosting extensions (Zhang and Yu, 2005), Bayesian extension of standard regression trees (Ročková and van der Pas, 2020), Bayesian extension of soft trees (Linero, 2018), to refer to tree-based methods. But one may also be interested in the uncertainty in the prediction, and thus consider prediction regions. For Lasso based estimators, Javanmard and Montanari (2014); van de Geer et al. (2014); Zhang and Zhang (2014) derive confidence regions for slope coefficient and statistical testing of sparsity for linear model using several tools: relaxed projection (Zhang and Zhang, 2014), desparsifying Lasso (van de Geer et al., 2014) or through the computation of an approximate inverse of the Gram matrix (Javanmard and Montanari, 2014). Since those pioneer works, several articles provide extensions for more general models or estimators, as generalised linear model (van de Geer et al. (2014) for convex loss function, Janková and van de Geer (2015) for subdifferential loss). We also refer to Meinshausen (2015) for groups of variables and Stucky and van de Geer (2018) for linear regression models with structured sparsity, among others. However, those results rely on strong assumptions on the design and although some authors consider more practical aspects (Chao et al., 2015; Lee et al., 2016), those results still remain difficult to be implemented.

We introduce in the following two contributions: we study the inverse regression in Section 1.1.1, for which we provide prediction regions, and we propose in Section 1.1.2 a new type of regression trees, that are smooth and can be used within each ensemble extension, and we derive their consistency.

1.1.1 Prediction regions through inverse regression

In this section, we propose to address the linear regression problem for elliptical distributions under an inverse regression approach rather than sparse regression. Introduced in Li (1991), the inverse regression relies on the Linearity Design Condition (LDC) which relates the covariates to elliptical distribution $\mathcal{E}_D(\mu, \Sigma, \phi)$ for a vector $\mu \in \mathbb{R}^d$, a positive semidefinite matrix $\Sigma \in \mathbb{R}^{d\times d}$ and a function $\phi : \mathbb{R}_+ \to \mathbb{R}$, characterized by the following theorem.

Theorem 1.1.1 (Cambanis et al. (1981))**.** *$X \sim \mathcal{E}_D(\mu, \Sigma, \phi)$ with $rank(\Sigma) = k$ if and only if*

$$X = \mu + \mathcal{R}\Lambda U^{(k)}$$

where the equality holds in distribution, and where $U^{(k)}$ is a k-dimensional random vector uniformly distributed on the unit hypersphere with $k-1$ dimensions $\mathcal{S}^{k-1}$, $\mathcal{R}$ is a non-negative random variable with distribution function F related to ϕ being stochastically independent of $U^{(k)}$, $\mu \in \mathbb{R}^D$ and $\Lambda \in \mathbb{R}^{D\times k}$ with rank$(\Lambda) = k$.

We propose to address the following linear regression problem with random regressors:

$$\mathbf{X}_i \sim \mathcal{E}_D(0, \mathbf{\Gamma}^\star, \phi) \text{ with } \text{rank}(\mathbf{\Gamma}^\star) = D \tag{1.1}$$
$$\mathbf{Y}_i|\mathbf{X}_i = \boldsymbol{A}^\star \mathbf{X}_i + \boldsymbol{\varepsilon}_i \tag{1.2}$$
$$\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_N \sim \mathcal{E}_L(0, \mathbf{\Sigma}^\star, \phi) \text{ with } \text{rank}(\mathbf{\Sigma}^\star) = L$$

where $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_N) \in \mathbb{R}^{L\times N}$ contains L responses for N subjects and $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_N) \in \mathbb{R}^{D\times N}$ contains D elliptical centered predictors with covariance matrix $\mathbf{\Gamma}^\star$. An interesting and relatively simple approach to handle the high dimensional problem, when D is large or/and when the number of observations N is smaller than D, is to consider the *inverse regression* problem:

$$\mathbf{Y}_i \sim \mathcal{E}_L(0, \mathbf{\Gamma}, \phi) \text{ with } \text{rank}(\mathbf{\Gamma}) = L \tag{1.3}$$
$$\mathbf{X}_i|\mathbf{Y}_i = \boldsymbol{A}\mathbf{Y}_i + \mathbf{e}_i \tag{1.4}$$
$$\mathbf{e} = (\mathbf{e}_1, \dots, \mathbf{e}_N) \sim \mathcal{E}_L(0, \mathbf{\Sigma}, \phi) \text{ with } \text{rank}(\mathbf{\Sigma}) = D$$

where $\boldsymbol{A}$ is a $D \times L$ matrix of slope coefficients of the *inverse regression*.

Interestingly, forward parameters $(\mathbf{\Gamma}^\star, \boldsymbol{A}^\star, \mathbf{\Sigma}^\star)$ are expressed in function of the inverse parameters $(\mathbf{\Gamma}, \boldsymbol{A}, \mathbf{\Sigma})$ through the following mapping:

$$\begin{aligned}\Psi : (\mathbf{\Gamma}, \boldsymbol{A}, \mathbf{\Sigma}) \mapsto & (\mathbf{\Gamma}^\star, \boldsymbol{A}^\star, \mathbf{\Sigma}^\star) \\ = & (\mathbf{\Sigma} + \boldsymbol{A}\mathbf{\Gamma}\boldsymbol{A}^\top, (\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\mathbf{\Sigma}^{-1}, (\mathbf{\Gamma}^{-1} + \boldsymbol{A}^\top\mathbf{\Sigma}^{-1}\boldsymbol{A})^{-1}).\end{aligned}$$

The mapping Ψ is an involution, the forward regression model (1.1)-(1.2) is then equivalent to the inverse regression model (1.3)-(1.4). The advantage of the inverse approach appears when structure is assumed on the large residual covariance matrix $\mathbf{\Sigma}$ in the inverse regression problem. Indeed, assuming that $\mathbf{\Sigma}$ is diagonal drastically reduces the number of parameters to estimate, while it implies a diagonal + low rank decomposition for $\mathbf{\Gamma}^\star$ through mapping Ψ: the residuals of the inverse model are not correlated while allowing structured correlations among covariates in the forward model. This is a strength of this model as in practice, correlated predictors often occur on real data. In this work, we study the uncertainty around the prediction

$$\widehat{\mathbf{Y}}_{N+1} = \mathbb{E}(\mathbf{Y}|\mathbf{X} = \mathbf{X}_{N+1}) = \widehat{\boldsymbol{A}}^\star \mathbf{x}_{N+1},$$

which can be quantified by deriving a prediction region. Assuming that covariance matrices $\mathbf{\Sigma}$ and $\mathbf{\Gamma}$ are both known and that $\mathbf{\Sigma}$ is diagonal as previously stated, we are able to derive the distribution of $\widehat{\boldsymbol{A}}^\star$.

Theorem 1.1.2 (Asymptotic distribution of $\widehat{\boldsymbol{A}}^\star$)**.** *Suppose $((\mathbf{X}_1, \mathbf{Y}_1), \dots, (\mathbf{X}_N, \mathbf{Y}_N))$ is a sequence of iid random variables satisfying the model defined in Equations* (1.1) *and* (1.2)*. Let*

$$\begin{aligned} g : \mathbb{R}^{D\times L} &\to \mathbb{R}^{L\times D} \\ \boldsymbol{A} &\mapsto \boldsymbol{A}^\star. \end{aligned}$$

Then, the following holds for the estimator $\widehat{\boldsymbol{A}}^\star$:

$$\sqrt{N}(vec(\widehat{\boldsymbol{A}}^\star) - vec(\boldsymbol{A}^\star)) \underset{N\to+\infty}{\to} \mathcal{N}_{DL}(\mathbf{0}, \Theta(\boldsymbol{A}));$$

where $\Theta(\boldsymbol{A}) = Cov(vec(Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A})))$. Moreover, $\Theta(\widehat{\boldsymbol{A}})$ is a consistent estimator of $\Theta(\boldsymbol{A})$, and

$$\sqrt{N}(vec(\widehat{\boldsymbol{A}}^\star) - vec(\boldsymbol{A}^\star))^T\Theta(\widehat{\boldsymbol{A}})^{-1}(vec(\widehat{\boldsymbol{A}}^\star) - vec(\boldsymbol{A}^\star)) \underset{N\to+\infty}{\to} \chi^2_{DL}.$$

where F is the distribution function of the random variable $\mathcal{R}$ involved in the stochastic representation of $vec(\boldsymbol{A}^\star)$, see Theorem 1.1.1.

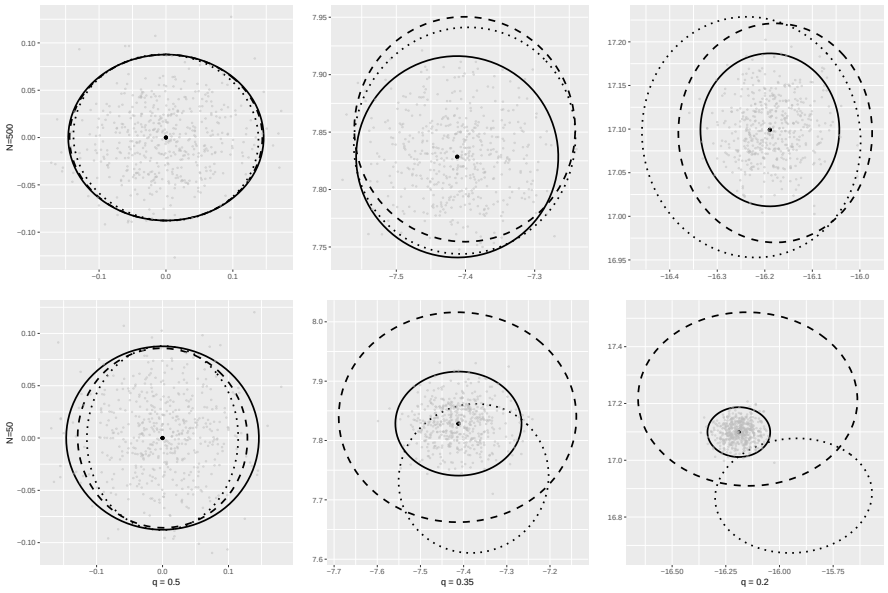


Figure 1.1: Prediction regions for $L = 2$. Dotted line: LSE for $N = 500$ and Bootstrapped Lasso for $N = 50$, long dashed line: IR, solid line: true parameters, grey dots: 1000 responses generated from the same covariate's profile.

This result provides closed-form expressions to derive confidence regions for $\boldsymbol{A}^\star$ and prediction regions.

Then, we provide the prediction region for a new observation.

Theorem 1.1.3. *Suppose* $((\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_N, \mathbf{Y}_N))$ *is a sequence of iid random variables satisfying the model defined in Equations* (1.1) *and* (1.2)*. Then, for* $\mathbf{X}_{N+1} \in \mathbb{R}^D$*,*

$$P\left(\mathbf{Y}_{N+1} \in \widetilde{\mathcal{PR}}_{\mathbf{Y},\alpha}\right) \underset{n\to+\infty}{\to} 1-\alpha$$

where

$$\widetilde{\mathcal{PR}}_{\mathbf{Y},\alpha} = \Big\{ y \in \mathbb{R}^L \; s.t. \qquad (1.5)$$
$$(y - \widehat{\boldsymbol{A}}^\star \mathbf{X}_{N+1})^T (\Omega(\boldsymbol{A}^\star \mathbf{X}_{N+1}) + \boldsymbol{\Sigma}^\star)^{-1} (y - \widehat{\boldsymbol{A}}^\star \mathbf{X}_{N+1}) \leq \chi^2_L(1-\alpha) \Big\}$$

where $\Omega(\boldsymbol{A}^\star \mathbf{X}_{N+1}) = (\mathbb{I}_L \otimes \mathbf{X}_{N+1}^T)\Theta(\boldsymbol{A})(\mathbf{X}_{N+1}^T \otimes \mathbb{I}_L)$*., where* $\Theta(\boldsymbol{A}) = Cov(vec(Dg(\boldsymbol{A}).(\widehat{\boldsymbol{A}} - \boldsymbol{A})))$*.*

One can notice that the covariance matrix that is inverted in Equation (1.5) breaks down into 2 parts. The first one, $\Omega(\boldsymbol{A}^\star \mathbf{X}_{N+1})$, represents the variance of the prediction which depends on the estimation accuracy of $\boldsymbol{A}^\star$ while the second part, $\boldsymbol{\Sigma}^\star$, is the variance inherited from the residuals.

Experimental validation

We consider a Gaussian setting, a response with dimension $L = 2$, $D = 100$ covariates, and $N \in \{50, 500\}$. We focus on sparse regression coefficients and independent responses: $\boldsymbol{A}$ is a $D \times L$ matrix with 90% of zero entries randomly drawn. The 10% nonzero remaining coefficients are uniformly drawn into a uniform distribution on $(-2, 2)$. Matrix $\boldsymbol{\Gamma}$ of covariances between response terms is set to $\mathbb{I}_L$. The residual covariance matrix of inverse regression $\boldsymbol{\Sigma}$ is set to $\mathbb{I}_D$. Note that a diagonal $\boldsymbol{\Sigma}$ and a sparse $\boldsymbol{A}$ under the inverse model lead to a sparse matrix of regression coefficients for forward regression $\boldsymbol{A}^\star$. For each simulated design, 1 000 learning datasets with dimension (N, D) are generated as well as 1 000 corresponding testing observations.

We compare the prediction regions derived from the 3 following methods: the proposed method based on inverse regression referred as IR in the following, the so-called least square estimator (LSE) for designs with $N > D$ and a lasso prediction interval based on bootstrap for designs with $N < D$. In this simulation study, the level of confidence for prediction regions is set to 95%. Figure 1.1 displays a graphical representation of prediction regions. Dotted line represents ellipses computed by LSE when $N = 500$ and Lasso when $N = 50$, long dashed line represents ellipses computed by IR and solid line represents true prediction regions computed with true parameters used for simulation. Grey dots are 500 replications of responses from the same covariate's profile representing the residual variance. Three specific profiles of covariates are considered: on the left panel, prediction ellipse for the median covariate's profile is computed which is an easy situation. When $N = 500$, both LSE and IR provide similar ellipses, close to the true one. When $N = 50$, IR's ellipse is close to the true one while lasso correctly predicts the response but the volume of the ellipse is larger. For the middle panel, a covariate's profile corresponding to quantile 0.35 is generated making the computation of the prediction ellipse more complex. When sample size is large, LSE and IR are competitive regarding to true ellipse and equivalent. When $N = 50$, the ellipse computed with IR is larger than the theoretical one. The bootstrapped Lasso fails in prediction. At last, for the right panel, an even more extreme profile associated to quantile 0.2 is generated, making the computation less reliable. When $N = 500$, the volume of ellipses computed by LSE and IR gets even larger as the covariate's profile gets far from the mean. Notice that LSE and IR again achieve similar ellipses in this setting. When $N = 50$, conclusions of the middle panel apply as well.

1.1.2 Probabilistic regression trees and their ensemble extensions

Let $Y \in \mathbb{R}$ be an output random variable linked to $\mathbf{X} = (X_1, \cdots, X_D)$ a D-dimensional input random vector through the following additive noise model:

$$Y = f(\mathbf{X}; \Theta) + \varepsilon_Y,\ \varepsilon_Y \sim \mathcal{N}(0, \tilde{\sigma}^2),$$

where Θ is the set of parameters on which f relies. In this section, we focus on a nonparametric method for regression, based on regression trees, that we call probabilistic regression trees (PR trees). We also extend it to ensemble methods, namely bagging, boosting and bayesian additive regression trees.

PR tree

The standard regression tree partitions the feature space into hyper-rectangles $(\mathcal{R}_k)_{1\leq k\leq K}$, referred to as regions, and assigns a weight γ_k to each region k:

$$f(\mathbf{x}; \Theta) = \sum_{k=1}^{K} \gamma_k \mathbf{1}_{\{\mathbf{x}\in\mathcal{R}_k\}},$$

where $\Theta = ((\mathcal{R}_k, \gamma_k)_{1\leq k\leq K})$. The splitting process, being dyadic, can be represented as a binary tree, where each node determines the features to split on and its corresponding value, resulting in the final partition given by the leaves of the tree. PR trees replace the indicator function from the standard regression trees with a function Ψ: for $\mathbf{x} \in \mathbb{R}^D$,

$$f_{\text{PR}}(\mathbf{x}; \Theta) = \sum_{k=1}^{K} \gamma_k \Psi(\mathbf{x}; \mathcal{R}_k, \boldsymbol{\sigma}).$$

The set of parameters to be estimated is thus $\Theta = ((\mathcal{R}_k)_{1\leq k\leq K}, \boldsymbol{\gamma}, \boldsymbol{\sigma})$, with $\boldsymbol{\sigma} \in \mathbb{R}_+^D$.

Given a training set $\mathcal{D}_N = \{(\mathbf{x}_i, y_i)_{1\leq i\leq N}\}$, for a fixed $\boldsymbol{\sigma}$, the regions and weights are updated until reaching a stopping criterion (weight estimate corresponds to the regression coefficient between $(y_i)_{1\leq i\leq N}$ and $\mathbf{P}$). During this process, the number of regions increases, and the matrix $\boldsymbol{P}$ and weights $\boldsymbol{\gamma}$ are gradually updated. Lastly, the vector $\boldsymbol{\sigma}$ can either be based on *a priori* knowledge or be learned through a grid search on a validation set. We rely on the latter in our experiments. Under some assumptions on the regularity of the true function (in a Sobolev

space), with number of regions going to infinity but not so fast with respect to the sample size and with diameter converging to 0, we derive the consistency of the parameter $\hat{\Theta}_N$ of PR tree:

$$\lim_{N\to+\infty} \mathbb{E}[|f_{\text{PR}}(\mathbf{X};\hat{\Theta}_N) - \mathbb{E}(Y|\mathbf{X})|^2] = 0.$$

Bagging and boosting ensemble methods

Bagging consists of averaging simple, noisy, but unbiased models. The random forest (RF) is a substantial modification of bagging, which builds a collection of decorrelated trees and then averages them. The extension of PR trees to RFs over m trees, denoted by PR-RF, is defined as follows:

$$f^{(m)}_{\text{PR-RF}}(\mathbf{x};\boldsymbol{\Theta}) = \frac{1}{m}\sum_{\ell=1}^{m} f_{\text{PR}}\left(\mathbf{x};\Theta^{(\ell)}\right)$$

where $\boldsymbol{\Theta} = (\Theta^{(1)},\ldots,\Theta^{(m)})$ with $\Theta^{(\ell)}$ characterizing the ℓ^{th} RF's tree in terms of parameters (i.e., split variables, cut points, predictions, and variances).

The standard RF is known to reduce the variance, because it averages identically distributed random variables (each tree), which are constructed to be the least correlated using bootstrap variables for each tree. This conclusion applies to any bagging extension, and specifically to the PR-RF, with the bias-variance trade-off being illustrated in the experiments. From a theoretical viewpoint, we can easily adapt the consistency result from Scornet et al. (2015) to derive the consistency of PR-RF when the number of trees m grows to infinity:

$$\lim_{N\to+\infty,m\to+\infty} \mathbb{E}[|f^{(m)}_{\text{PR-RF}}(\mathbf{X};\hat{\boldsymbol{\Theta}}_N) - \mathbb{E}(Y|\mathbf{X})|^2] = 0.$$

On the other hand, boosting methods gradually improve the prediction by optimizing the residuals with respect to the prediction based on the model constructed so far. Let us assume that $(m-1)$ PR trees have been built so far. The m^{th} smooth tree and its parameter $\hat{\Theta}^{(m)}$, defined by

$$\hat{\Theta}^{(m)} := \underset{\Theta^{(m)}}{\operatorname{argmin}} \sum_{i=1}^{N}\left(\left(y_i - \sum_{\ell=1}^{m-1}\sum_{k=1}^{K^{(\ell)}} \gamma_k^{(\ell)}[P_{ik}]^{(\ell)}\right) - \sum_{k=1}^{K^{(m)}} \gamma_k^{(m)}[P_{ik}^{(m)}]\right)^2$$

where the matrices $\mathbf{P} = (P_{ik}^{(m)})_{1\le i\le N, 1\le k\le K^{(m)}}$ depend on the regions. So, the prediction function is given by:

$$f^{(m)}_{\text{PR-GBT}}(\mathbf{x};\boldsymbol{\Theta}) = \sum_{\ell=1}^{m} f_{\text{PR}}\left(\mathbf{x};\Theta^{[\ell]}\right)$$

where $\boldsymbol{\Theta} = (\Theta^{[1]},\ldots,\Theta^{[m]})$.

Boosting methods are known to reduce the bias of the prediction function (while allowing for a small variance), which is true for our PR-GBT prediction function, as illustrated in the experiments (see Figure 1.2). However, boosting forever can overfit the data, and it is thus necessary to stop the procedure with an adaptive finite number of steps (Zhang and Yu, 2005). In theory, we mimic the results obtained in Zhang and Yu (2005) and so we apply early stopping; in practice, we fix the number of trees. Note that this context is more difficult in theory, because each tree is dependent on the previous one with a random number of trees. We achieve convergence in probability instead of an L_2 convergence:

$$f^{\hat{m}}_{\text{PR-GBT}}\left(\mathbf{X};\hat{\boldsymbol{\Theta}}_N\right) \xrightarrow[N\to+\infty]{\mathbf{P}} \mathbb{E}[Y|\mathbf{X}].$$

Probabilistic Bayesian additive regression trees

The Bayesian additive regression tree is a boosting extension, where *a priori* distribution are settled to add randomness. In this section, we describe how to construct P-BART, an extension of BART (Chipman et al., 2010) using PR trees.

We consider m distinct regression trees, with the ℓ^{th} tree having a tree structure $T^{(\ell)}$ and weights $\boldsymbol{\gamma}^{(\ell)} = (\gamma_1^{(\ell)}, \ldots, \gamma_K^{(\ell)})$. Trees are fitted iteratively until no change is observed, thus holding all other $m-1$ trees constant and considering the residual response that remains unfitted. The prediction using P-BART is made as an averaging over the iterations (after burning) of the sum (boosting part) of the prediction for a new covariate $\mathbf{x}$ through a PR tree.

$$f_{\text{P-BART}}(\mathbf{x};\Theta) = \frac{1}{\text{it} - \text{it}_{\text{burn}}} \sum_{t=\text{it}_{\text{burn}+1}}^{\text{it}} \sum_{\ell=1}^{m} \sum_{k=1}^{K^{(\ell)}} \gamma_k^{(\ell),t} \Psi(\mathbf{x}; \mathcal{R}_k^{(\ell),t}, \boldsymbol{\sigma}),$$

where Θ corresponds to all the parameters needed to define P-BART, $T^{(\ell),t}$ is the ℓth tree at iteration t, and similarly its parameters, while it and it_{brun} are respectively the number of iterations performed and the number of iterations for the burning.

We derive the prior and posterior distribution associated to our PR trees for the bayesian additive context, details are omitted here.

Under classical assumptions for BART (detailed in the main paper) and similar assumptions as before for probabilistic regression tree, we prove the convergence of the posterior distribution to the true function, defined by

$$\Pi_N(A) = \frac{\int_A \prod_{i=1}^N p_f(y_i|\mathbf{x}_i)\Pi(df)}{\int \prod_{i=1}^N p_f(y_i|\mathbf{x}_i)\Pi(df)},$$

where Π denotes the prior probability measure over $L^2([0,1]^D)$: for $N\varepsilon_N^2 \to \infty$ and $\varepsilon_N \to 0$ as $N \to \infty$, we get, for some $M_1 > 0$,

$$\Pi_N\left(\|\widetilde{f} - f\|_N \geq M_1\varepsilon_N^2\right) \xrightarrow[N\to+\infty]{\mathbf{P}} 0.$$

Experimental validation

Detailed results are provided in the main paper, but we illustrate here the evolution of the bias and variance for each estimator on the Diabetes data set. For the prediction function $\hat{f}$, we define:

$$\text{bias}(\hat{f}) = \text{E}_{(\mathbf{X},Y)}\left\{\left(Y - \text{E}\{\hat{f}(\mathbf{X})\}\right)^2\right\};$$
$$\text{var}(\hat{f}) = \text{E}_{(\mathbf{X},Y)}\left(\text{Var}\{\hat{f}(\mathbf{X})\}\right);$$

where the inner expectation and variance are with respect to the estimator. To compute the bias and variance, we subsample the data with 80% for training and 20% for testing, which estimates the inner and outer expectations.

In Fig. 1.2, we provide the result for one tree, RF, GBT, and BART. It is well known that bagging improves the variance (and makes it decrease with the number of trees). This is indeed illustrated in the plots. However, we observe that standard RF, PR-RF, and STR-RF achieve the same variance. At the end, PR-RF has the best performance in RMSE, because it improves (even for one tree) the bias. Note that the plot of bias is very similar for one tree and RF. It is also known that boosting reduces the bias, as illustrated in this figure. Again, all methods perform similarly, and the gain for PR-RF in RMSE is achieved thanks to the variance reduction for one tree. Finally, as BART is a mixture between bagging and boosting (after a warming phase), we recognize the improvement in both the bias and variance. All the methods provide comparable results, the standard BART having slightly worse performance and PR-BART a slightly better performance with 100 trees.

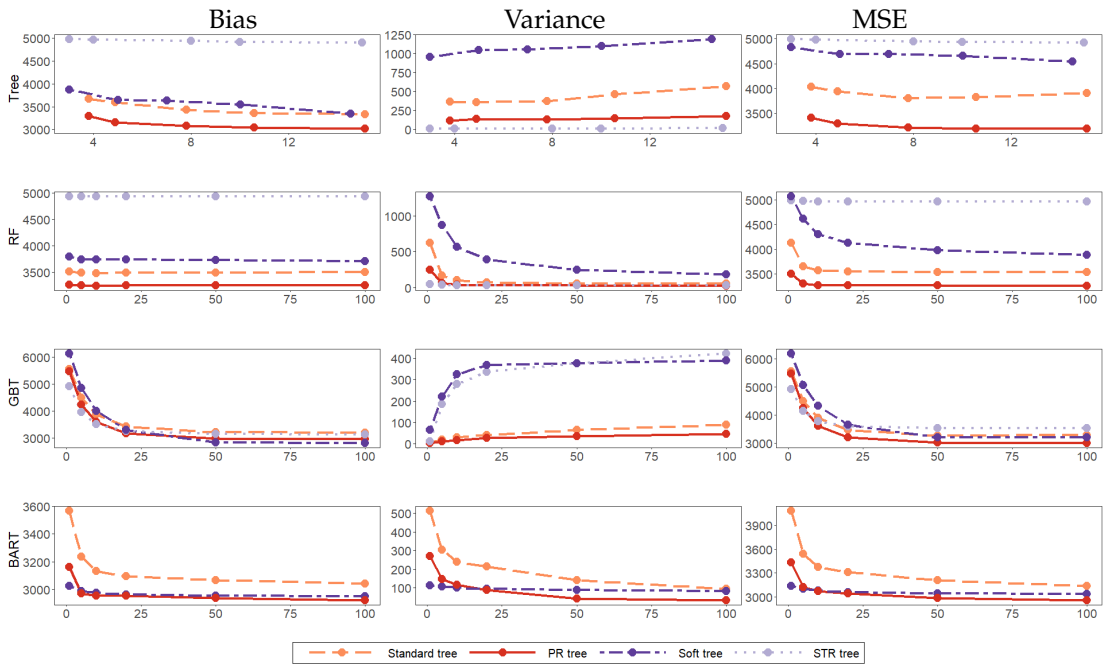


Figure 1.2: Evolution of the performance of one tree (first row), bagging methods (second row), boosting methods (third row), and Bayesian ensemble methods (fourth row) for the Diabetes data set. Left: bias, middle: variance, right: MSE. For one tree, we increase the depth of the tree to vary the dimension, whereas for the ensemble methods, we increase the number of trees.

1.2 Functional data

Functional Data Analysis (FDA) deals with the theory and the exploration of data observed over a finite discrete grid and expressed as curves (or mathematical functions) varying over some continuum such as time (Ferraty and Vieu, 2006; Ramsay and Silverman, 2002, 2005).

When analyzing functional data, the questions are slightly different than the one from iid data. One can wonder: (i) is there a common main (mean) functional pattern to be distinguished?; (ii) can we quantify the significant individual fluctuations with respect to such a mean pattern?; (iii) can we disentangle the high heterogeneity of the data both at the level of the studied individuals and on the time dimension?; (iv) does the uncertainty around the estimation of the common main functional pattern can be measured?

We consider multiple independent observations of the same function on several timepoints, yielding noisy functional data $(Y_i(t_j))_{1\leq i\leq N, 1\leq j\leq n}$. In the several contributions presented in this manuscript, we consider different models, but to fix the idea we can start with the linear functional model:

$$Y_i(t_j) = f(t_j) + \varepsilon_{ij}, \tag{1.6}$$

where $\varepsilon_{i.} = (\varepsilon_{i1}, \ldots, \varepsilon_{in})$ is the noise representing the individual functional variation around f. We assume that the ε_i are independent.

To analyze such data, a common approach, typically in the parametric setting, involves projecting the data onto a functional space defined by a family of functions (Li et al., 2022; Kokoszka and Reimherr, 2017). We assume that the global mean f belongs to a functional space $\mathcal{S}^{L^*} = Vect((t \mapsto B_\ell^{L^*}(t))_{1\leq \ell\leq L^*})$ with L^* functions $(B_\ell^{L^*})_{1\leq \ell\leq L^*}$ assumed to be linearly independent, and can be written as $f(t) = f^{L^*}(t) = \sum_{\ell=1}^{L^*} \mu_\ell^{L^*} B_\ell^{L^*}(t) \in \mathcal{S}^{L^*}$. Several families can be considered, as the B-splines (considered in Section 1.2.1), wavelets (considered in Section 1.2.2). One can also expect more structure, through the following assumption.

Assumption 1.2.1. *The functional family $(t \mapsto B_\ell^{L^*}(t))_{1\leq \ell\leq L^*}$ is orthonormal with respect to the standard scalar product $< ., . >$.*

Note that if Assumption 1.2.1 holds, one get $\mu_\ell^{L^*} =< f^{L^*}, B_\ell^{L^*} >$ for $\ell = 1, \ldots, L^*$. The Legendre family is orthonormal, the Fourier family is orthogonal for the standard scalar product (but not orthonormal), and the B-splines family is not orthogonal.

Depending on the modelisation, we can assume that the noise is a white noise, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I)$, or that there exists a sequence of coefficients $(c_{i\ell})_{1\leq \ell\leq L^\varepsilon}$ such that $\varepsilon_{ij} = \sum_{\ell=1}^{L^\varepsilon} c_{i\ell} B_\ell^{L^\varepsilon}(t_j)$, with $c_{i\ell} \sim_{iid} \mathcal{N}(0, \sigma^2)$.

Let $f^{L^*} \in \mathcal{S}^{L^*}$ with L^* unknown, and consider the space $\mathcal{S}^L$ for L fixed. As $\mathcal{S}^L$ is a family of linearly independent functions, there always exists a unique vector μ^{L,L^*} of coefficients defining $f^{L,L^*}(t) = \sum_{\ell=1}^{L} \mu_\ell^{L,L^*} B_\ell^L(t) = B^L(t)\mu^{L,L^*}$ such that

$$f^{L,L^*} = \arg\min_{f\in\mathcal{S}^L} \{\|f^{L^*} - f\|_2^2\},$$

and if the family is orthonormal (Assumption 1.2.1), it corresponds to the projected coefficients $\mu_\ell^{L,L^*} :=< f^{L^*}, B_\ell^L >$. We can prove that if the family is orthonormal, $f^{L^*,L^*} = f^{L^*}$ and the projection coefficients verify $\mu_\ell^{L,L^*} = \mu_\ell^{L^*}$ for $\ell = 1, \ldots, \min(L, L^*)$.

As the observations are recorded at discrete time points $(t_j)_{1\leq i\leq N, 1\leq j\leq n}$, we introduce the family of functions evaluated at the discrete times of observations. For $L \in \mathbb{N}$, let us denote $\mathbf{B}^L$ the matrix of $n \times L$ with coefficient in row j and column ℓ equal to $B_\ell^L(t_j)$. We consider the operator $\mathbf{P}^L$ defined as the matrix $\mathbf{P}^L = ((\mathbf{B}^L)^T\mathbf{B}^L)^{-1}(\mathbf{B}^L)^T$ of size $L \times n$ (this operator is a bit more complex when the functional family is not orthonormal wrt the standard scalar product). Then we define the coefficients $\underline{\mu}^{L,L^*}$ which are the coefficients of μ^{L,L^*} approximated on the vector space, denoted $\mathbf{S}^L$, defined by the matrix $\mathbf{B}^L$.

$$\underline{\mu}^{L,L^*} := \mathbf{P}^L\mathbf{B}^{L^*}\mu^{L^*}.$$

The corresponding finite approximated regression function is denoted $\underline{f}^{L,L^*}$ and is defined, for all $t \in [0, 1]$, as

$$\underline{f}^{L,L^*}(t) = B^L(t)\underline{\mu}^{L,L^*}.$$

When considering the estimation of the regression function f^{L^*} on the space of dimension L defined by the family $\mathbf{B}^L$, we do not directly estimate f^{L^*} but its projection on this finite space, which corresponds to the projected function $\underline{f}^{L,L^*}(t)$ and its associated coefficients $(\underline{\mu}_\ell^{L,L^*})_{1\leq\ell\leq L}$.

Proposition 1.2.2. *The vector of coefficients $(\underline{\mu}_\ell^{L,L^*})_{1\leq\ell\leq L}$ is estimated by the least square estimator $\underline{\hat{\mu}}^{L,L^*}$ defined as:*

$$\underline{\hat{\mu}}^{L,L^*} := \frac{1}{N}\sum_{i=1}^{N}\mathbf{P}^L y_{i.} \sim \mathcal{N}_L\left(\underline{\mu}^{L,L^*}, \frac{\sigma^2}{N}\mathbf{P}^L\mathbf{B}^{L^\varepsilon}(\mathbf{B}^{L^\varepsilon})^T(\mathbf{P}^L)^T\right).$$

For a fixed $t \in [0,1]$, the estimator of the function $\underline{f}^{L,L^*}(t)$ is defined by:

$$\underline{\hat{f}}^{L,L^*}(t) = \sum_{\ell=1}^{L}\underline{\hat{\mu}}_\ell^{L,L^*}B_\ell^L(t) = B^L(t)\underline{\hat{\mu}}^{L,L^*}.$$

This comes back to the standard linear regression, and we consider the classical least square estimator. However, it induces a bias that will be complex to evaluate, as described in Section 1.2.3.

One can then consider more complex models to describe functional data and answer the first questions. To describe the **synchronization over time** of the curves, one can preprocess them or take it as a part of the modeling process. Aligning the individual curves via individual shift functions is conveniently done via time warping functions, see for example Bigot (2013); Claeskens et al. (2010); Dupuy et al. (2011); Gervini and Gasser (2004); Kneip and Gasser (1992); Wang and Gasser (1997). One approach towards describing the curve-specific deviations from the mean curve is via random effects, see for example Chen and Wang (2011); Elmi et al. (2011); Guo (2002). An analysis of variance model for functional data describing the phase variability through time-warping and allowing for inference in the presence of amplitude variability, was introduced by Gervini and Carter (2014). A functional mixed effects regression model was used to analyse spike train data in Hadjipantelis et al. (2014). A shift-warping method is used in Carroll et al. (2020) for multivariate functional data where each of the components may contribute to a shift with its own parameter value. A nonparametric registration method is proposed in Chakraborty and Panaretos (2021), based on a local variation measure introduced to provide nonparametric conditions that lead to identifiability. The phase and amplitude are separated in Tucker et al. (2013) by using a representation of functional data that is based on the Fisher-Rao metric to compute an elastic shape analysis of the curves. Based on this representation, Yu et al. (2017) analyses the phase variation using a principal nested sphere approach. In Strait et al. (2017), a constrained elastic shape analysis is used with a landmark representation. While there are Bayesian methods for registration too, see for example Cheng et al. (2016), these are not considered here. In Section 1.2.1, we introduce a new nonlinear mixed effects modeling and warping, and prove the consistency of the associated estimator.

Another modelisation aspect is the heterogeneity. **Model-based clustering approaches** for functional data have been extensively studied in the literature (James and Sugar, 2003; Liu and Yang, 2009; Bouveyron and Jacques, 2011; Jacques and Preda, 2013, 2014; Devijver, 2017b). For the particular case of heterogeneous data that interests us in this article, one can broadly differentiate between methods that perform simultaneously clustering and segmentation and co-clustering based methods.

Samé et al. (2011) proposed to deal with heterogeneous time series by integrating the notion of change of regimes within a mixture of hidden logistic process regressions. The model is considering two latent variables, one for the mixture component and one for the segmentation. Model selection is done through an adapted BIC criterion. However, while attempting to consider changes of regime, this approach fails to account for the ordering of observations, a key feature when dealing with functional data. Samé and Govaert (2012) extended this model for online segmentation of time series. In an effort to account for these potential changes of regimes, another family of mixture models, namely the mixture of piecewise regression, has been proposed. Hébrail et al. (2010) first define this notion of piecewise regression to analyze temporal data, by proposing a distance-based model that simultaneously performs clustering

on the set of functional observations (through a Kmeans-like algorithm) and segmentation (in the form of piecewise constant function summarizing) within each of the obtained cluster. This work was further generalized to a more flexible probabilistic framework by Chamroukhi (2016), who designed a model based on a mixture of piecewise regression densities. The piecewise regression is modeled by a segmentation of polynomial functions, as a generalization of spline basis where knots have to be fixed. However, this sets a particular form within each segment. Bouveyron et al. (2017) proposed a co-clustering model to analyze multivariate functional data. They apply this model to analyze electricity consumption curves, and found that due to the nature of the temporal data, the clustering over timepoints is in fact close to a segmentation over time. Bouveyron et al. (2021b) extend this method to multivariate time series (with several time series for each observation and each timepoint), using a sparse representation over principal components. In Bouveyron et al. (2021a), authors extend this co-clustering approach using a shape invariant model, allowing for translation in time, and translation and scaling in mean. Galvani et al. (2021) propose another bi-clustering algorithm for functional data while considering a potential misalignment through translation. While co-clustering based approach have proven efficient in this context, the clustering obtained on the time dimension do not account for the ordering of the observation. In Section 1.2.2, we introduce a new mixture of segmentation model, ad prove the consistency of the associated estimator.

Finally, providing **simultaneous confidence bands** for the function means, rather than pointwise confidence intervals, is essential to measure uncertainty around the global mean and extend Proposition 1.2.9. This task presents several challenges: the confidence band must control the simultaneous functional type-I error rate, balance being conservative enough to maintain confidence without being overly so, and be computationally feasible. Several methods have been proposed to address these issues. For datasets with many time points but no repetition, asymptotic distribution methods study the infinity norm between the true function and its estimator (Hall, 1991; Claeskens and Van Keilegom, 2003), though these can be overly conservative for smaller samples. Bootstrap methods, while suitable for small samples, are computationally intensive. The volume-of-tube formula approach, used by Sun and Loader (1994), Zhou et al. (1998), and Krivobokova et al. (2010), constructs confidence bands using an unbiased linear estimator. Wang et al. (2022) and Liebl and Reimherr (2023) further developed methods to reduce computational complexity and variability. Some studies, like Bunea et al. (2011) and Telschow et al. (2023), rely on multiple observations of the same function, proposing threshold-type estimators and bands based on the Gaussian kinematic formula. Recent extensions have been proposed, to nonstationary random field in Telschow et al. (2023), based on conformal prediction in Diquigiovanni et al. (2022), or having a prediction goal in mind in Nicolás Hernández and Jacques (2024) by considering functional time series data set. However, a common limitation is not accounting for the bias of the functional estimator. Sun and Loader (1994) proposed a bias correction, but it remains impractical due to the open choice of smoothing parameters. In nonparametric frameworks, bias is approximated using second derivative estimators, but general solutions for bias handling are scarce. Section 1.2.3 aims to address the bias issue in confidence band construction for general functions using a finite functional orthonormal family.

1.2.1 Nonlinear mixed effects modeling and warping for functional data

In this section, we propose a nonlinear functional mixed model to represent variability in amplitude and phase. We provide an estimator that is proven to be consistent, along with a convergent algorithm.

The model and its identifiability

Suppose one observes individual curves $Y_1(t), Y_2(t), \dots, Y_N(t)$ on the interval $[0,1]$ (without loss of generality), and a first aim is to find a main pattern $\mu(t)$ in these individual curves.

We consider the following discretization of the functional mixed model with time points $(t_{i,j})$ for $j = 1, \dots, T_i$; $i = 1, \dots, N$, where T_i denotes the number of fixed (non-random) time points for the individual i:

$$Y_i(t_{i,j}) = \mu\left\{w^{-1}(t_{i,j};\boldsymbol{\theta}_i)\right\} + U_i\left\{w^{-1}(t_{i,j};\boldsymbol{\theta}_i)\right\} + \varepsilon_{i,j}, \tag{1.7}$$

with μ the unknown common mean, U_i denotes the unknown random effect on the amplitude for the observation i, $w : [0,1] \to [0,1]$ is a flexible warping function strictly increasing and depending on a random variable $\boldsymbol{\theta}_i \sim \mathcal{N}_r(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$, that describes the individual phase variability. We assume that for all i, the error vectors $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,T_i})^\top$ with $\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{T_i}(\mathbf{0}_{T_i}, \sigma^2_{\boldsymbol{\varepsilon}} \mathbf{I}_{T_i})$ are i.i.d., meaning that the error terms are independent of $t_{i,j}$ and of the warping effects $\boldsymbol{\theta}_i$.

The warping function w, the unknown mean function μ and the individual random effect amplitude functions U_i are modeled in a flexible fashion via B-splines of degree p and with K interior knots $0 = \kappa_0 < \ldots < \kappa_K + 1 = 1$, such that: for every $t \in [0,1]$,

$$\begin{aligned}
\mu(t) &= \sum_{\ell=-p_\mu}^{K_\mu} \alpha^\mu_\ell B^\mu_{\ell,p_\mu+1}(t;\boldsymbol{\kappa}^\mu), \\
U_i(t) &= \sum_{l=-p_{U_i}}^{K_{U_i}} \alpha^U_{i,l} B^U_{i,l,p_{U_i}+1}(t;\boldsymbol{\kappa}^{U_i}), \\
w^{-1}(t;\boldsymbol{\theta}_i) &= \frac{\int_0^t \exp\left\{h^{-1}(u;\boldsymbol{\theta}_i)\right\} du}{\int_0^1 \exp\left\{h^{-1}(u;\boldsymbol{\theta}_i)\right\} du}, \\
h^{-1}(u;\boldsymbol{\theta}_i) &= \sum_{l=-p_h}^{K_h} \theta_{i,l} \bar{B}^h_{l,p_h+1}(u;\boldsymbol{\kappa}^h),
\end{aligned}$$

where $\boldsymbol{\alpha}^U_i = (\alpha^U_{i,-p_{U_i}}, \ldots, \alpha^U_{i,K_{U_i}})^\top$ is such that

$$\boldsymbol{\alpha}^U_i = \begin{pmatrix} \alpha^U_{i,-p_{U_i}} \\ \vdots \\ \alpha^U_{i,K_{U_i}} \end{pmatrix} \sim \mathcal{N}_{m_{U_i}}\left(\mathbf{0}_{m_{U_i}}, \Sigma^{U_i}\right) \text{ with } \Sigma^{U_i} = \begin{pmatrix} \sigma^2_{U,1} & 0 & \ldots & 0 \\ 0 & \sigma^2_{U,2} & \ldots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \ldots & 0 & \sigma^2_{U,m_{U_i}} \end{pmatrix},$$

the covariance matrix for which we assume a diagonal structure, and which needs to be estimated. Further we denote $\underset{\sim}{\boldsymbol{\alpha}}^U = \left((\boldsymbol{\alpha}^U_1)^\top, \ldots, (\boldsymbol{\alpha}^U_N)^\top\right)^\top$, a random vector taking values in $\mathbb{R}^{\sum_{i=1}^N m_{U_i} \times 1}$. Note that w^{-1} (and hence w) is by construction an increasing function. A non-random version of this warping function was introduced in Ramsay and Silverman (2005). To ensure identifiability, the function h^{-1} will be decomposed using a basis of centralized B-splines, i.e. we consider $(\bar{B}^h_{l,p_h+1})_{l=-p_h,\ldots,K_h}$ satisfy

$$\int_0^1 \bar{B}^h_{l,p_h+1}(u;\boldsymbol{\kappa}^h) du = 0.$$

The vector of random effects $\boldsymbol{\theta}_i = (\theta_{i,-p_h}, \ldots, \theta_{i,K_h})^\top$ describes the individual phase variability, for which we assume a linear mixed effects model

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \mathbf{E}_i + \tilde{\boldsymbol{\varepsilon}}_i,$$

with $\mathbf{E}_i \sim \mathcal{N}_r(\mathbf{0}_r, \Sigma^{\mathbf{E}})$ and $\tilde{\boldsymbol{\varepsilon}}_i \sim \mathcal{N}_r(\mathbf{0}_r, \sigma^2_{\tilde{\boldsymbol{\varepsilon}}} \mathbf{I}_r)$ independent. Then $\boldsymbol{\theta}_i \sim \mathcal{N}_r(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$, with $\Sigma^{\boldsymbol{\theta}} = \Sigma^{\mathbf{E}} + \sigma^2_{\tilde{\boldsymbol{\varepsilon}}} \mathbf{I}_r$. To ensure identifiability, we assume that $\sigma^2_{\tilde{\boldsymbol{\varepsilon}}}$ is known. Further, we assume that the $\boldsymbol{\alpha}^U_i$s and $\boldsymbol{\theta}_i$s, the random effects describing the individual phase and amplitude variability, are independent of each other.

For further analysis it will be useful to introduce some matrix notation. The matrix $\mathbf{B}^\mu_i$ of dimension $T_i \times m_\mu$ contains (j,l)th element $B^\mu_{l,p_\mu+1}(t_{ij};\boldsymbol{\kappa}^\mu)$, and $\mathbf{B}^U_i$ is the matrix of dimension $T_i \times m_{U_i}$ with (j,l)th element $B^U_{i,l,p_{U_i}+1}(t_{ij};\boldsymbol{\kappa}^{U_i})$. Further, $(\mathbf{B}^\mu_i)^{\boldsymbol{\theta}_i} = \left([(\mathbf{B}^\mu_i)^{\boldsymbol{\theta}_i}]_{j,l}\right)_{j=1,\ldots,T_i; l=-p_\mu,\ldots,K_\mu}$, with

$$[(\mathbf{B}^\mu_i)^{\boldsymbol{\theta}_i}]_{j,l} = B^\mu_{l,p_\mu+1}\{w^{-1}(t_{i,j};\boldsymbol{\theta}_i);\boldsymbol{\kappa}^\mu\}.$$

Define $[(\mathbf{B}^\mu_i)^{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}]_{j,l} = B^\mu_{l,p_\mu+1}[w^{-1}\{w(t_{i,j};\boldsymbol{\theta}_i);\tilde{\boldsymbol{\theta}}_i\};\boldsymbol{\kappa}^\mu]$ for $j = 1,\ldots,T_i$ and $l = -p_\mu,\ldots,K_\mu$. Similarly, we define $(\mathbf{B}^U_i)^{\boldsymbol{\theta}_i}$ and $[(\mathbf{B}^U_i)^{\boldsymbol{\theta}_i,\tilde{\boldsymbol{\theta}}_i}]_{j,l}$.

Finally, we provide sufficient and necessary conditions to ensure the identifiability of model. First, note that it is identifiable if and only if at least one (approximate) individual model (1.7) is identifiable. We thus focus on a fixed i, and on the set of parameters $(\boldsymbol{\alpha}^\mu, \sigma_\varepsilon^2, \Sigma^{U_i}, \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$. Remark also that if we know $(\sigma_\varepsilon^2, \Sigma^{U_i})$, or if we know σ_ε^2, or if we know Σ^{U_i}, the problem is much simpler.

Theorem 1.2.3. *Let $i \in \{1, \dots, n\}$ be given. Let $\boldsymbol{\theta}_i \sim \mathcal{N}_r(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$ and $\tilde{\boldsymbol{\theta}}_i \sim \mathcal{N}_r(\tilde{\boldsymbol{\theta}}_0, \Sigma^{\tilde{\boldsymbol{\theta}}})$ be used to define two warping functions $w^{-1}(.;\boldsymbol{\theta}_i)$ and $w^{-1}(.;\tilde{\boldsymbol{\theta}}_i)$, and let X_i and $\tilde{X}_i$ be the corresponding warped functions, such that*

$$Y_i(t) = X_i\{w^{-1}(t;\boldsymbol{\theta}_i)\} = \tilde{X}_i\{w^{-1}(t;\tilde{\boldsymbol{\theta}}_i)\}.$$

Then model (1.7) *is identifiable if and only if*

$$\begin{aligned}
\mathbf{B}_i^\mu &= \mathrm{E}_{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i}\left\{(\mathbf{B}_i^\mu)^{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i}\right\}; \\
(\mathbf{B}_i^U)^\top \Sigma^{U_i} \mathbf{B}_i^U &= \mathrm{Var}_{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i}\left\{(\mathbf{B}_i^\mu)^{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i}\boldsymbol{\alpha}^\mu\right\} + \mathrm{E}_{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i}\left[\left\{(\mathbf{B}_i^U)^{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i}\right\}^\top \Sigma^{U_i} (\mathbf{B}_i^U)^{\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i}\right].
\end{aligned}$$

and at least one of the three condition is not satisfied:

1. $(\mathbf{B}_i^U)^\top \mathbf{B}_i^U \neq \mathbf{0}_{m_{U_i}}$;
2. $\mathbf{H}_i^U = \mathbf{B}_i^U \{(\mathbf{B}_i^U)^\top \mathbf{B}_i^U\}^{-1} (\mathbf{B}_i^U)^\top = \mathbf{I}_{T_i}$;
3. $(\mathbf{B}_i^U)^\top \mathbf{B}_i^U$ *is diagonal.*

The proof relies on the iteration of two identifiable steps until convergence, the one for warping parameters and the one of the warped process. Note that the identifiability conditions are essentially conditions on the englobing B-spline basis structure.

Estimation procedure and asymptotic properties

Recall that the unknown parameters of model (1.7) are $(\boldsymbol{\alpha}^\mu, \sigma_\varepsilon^2, \underset{\sim}{\Sigma}^U, \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$. Model (1.7) is a *nonlinear* functional mixed effects model due to the composition by the warping function, which is an essential ingredient to describe the individual phase variability. First, we analyse each part of the model, that is, the warping parameters and the linear mixed effect model, by providing an estimator and theoretical guarantees. Then, we propose an iterative estimation procedure, where in a first step we fix the warping parameters $(\boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}})$ and estimate the functional parameters $(\boldsymbol{\alpha}^\mu, \sigma_\varepsilon^2, \underset{\sim}{\Sigma}^U)$; and next, we start from these estimated parameters, and estimate the warping parameters. In Theorem 1.2.4 we prove that the algorithm is converging.

We denote by $((\hat{\boldsymbol{\alpha}}^\mu)^{(\infty)}, (\hat{\sigma}_\varepsilon)^{(\infty)}, (\underset{\sim}{\hat{\Sigma}}^U)^{(\infty)}, \hat{\boldsymbol{\theta}}_0^{(\infty)}, (\hat{\Sigma}^{\boldsymbol{\theta}})^{(\infty)})$ the estimator obtained at the end of the algorithm. First we derive the pointwise convergence of the algorithm.

Theorem 1.2.4. *Fix N and $\mathbf{T}$. Suppose $(\mathbf{Y}_1, \dots, \mathbf{Y}_N)$ is a sequence of iid random variables satisfying the functional nonlinear mixed model* (1.7) *observed on fixed time points: for $i = 1, \dots, N$, for $j = 1, \dots, T_i$, $[\mathbf{Y}_i]_j = Y_i(t_{i,j})$. Moreover, suppose that the model is identifiable and that the update of $\boldsymbol{\theta}_i$ is a contraction mapping.*

Then, $((\hat{\boldsymbol{\alpha}}^\mu)^{(\infty)}, (\hat{\sigma}_\varepsilon)^{(\infty)}, (\underset{\sim}{\hat{\Sigma}}^U)^{(\infty)}, \hat{\boldsymbol{\theta}}_0^{(\infty)}, (\hat{\Sigma}^{\boldsymbol{\theta}})^{(\infty)})$ exists and is unique, and the algorithm converges to this solution with a geometric rate with respect to the Euclidean distance.

As we are working with a nonlinear least square estimator, we need to define the weighted tail product, first introduced in Jennrich (1969).

Definition 1.2.5. *Let p be a nonnegative integer and $(t_j)_{j=1,\dots,p}$ be fixed time points. Let $x = (x_p)$ and $y = (y_p)$ be two sequences of real numbers and let*

$$(x, y)_p^\pi = \frac{1}{p} \sum_{j=1}^{p-1} x_j y_j (t_{j+1} - t_j).$$

If $(x,y)_p^\pi$ converges to a real number when $p \to +\infty$, its limit $(x,y)^\pi$ is called the weighted tail product of x and y. For $l = 1,\dots,r$, we denote by

$$\partial_l(\mu + U_i) = \frac{\partial[(\mu(w^{-1}(t_{i,j};\tilde{\boldsymbol{\theta}}_i)) + U_i(w^{-1}(t_{i,j};\tilde{\boldsymbol{\theta}}_i)))_{j=1,\dots,T_i}]}{\partial[\tilde{\boldsymbol{\theta}}_i]_l}.$$

the partial derivative of the aligned signal. We define

$$\mathbf{a}_{i,T_i}(\tilde{\boldsymbol{\theta}}_i) = \left[(\partial_l(\mu + U_i), \partial_{l'}(\mu + U_i))_{T_i}^\pi\right]_{l=1,\dots,r;l'=1,\dots,r}$$

the matrix with coefficients the weighted tail product between two partial derivatives, and $\mathbf{a}_i(\tilde{\boldsymbol{\theta}}_i)$ its limit when $T_i \to +\infty$.

Suppose we know the warping parameters $(\boldsymbol{\theta}_i)_{i=1,\dots,N}$. Then, we warp the observations $(Y_i(t_{i,j}))_{j=1,\dots,T_i;i=1,\dots,N}$ onto the estimated warped curves $X_{i,j} = Y_i\{w(t_{i,j};\boldsymbol{\theta}_i\}$, and we fit a functional linear mixed model on $(\mathbf{X}_i)_{i=1,\dots,N}$ using maximum likelihood estimation, which leads to estimators $(\hat{\boldsymbol{\alpha}}^\mu, \hat{\Sigma}^U, (\hat{\sigma}_\varepsilon^2))$ and predictors $(\hat{\boldsymbol{\alpha}}_i^U)_{i=1,\dots,N}$.

We finally provide the statistical consistency of the full procedure. This has the following meaning. When the sample size and the number of time points are going to infinity, the parameters estimated by the iterative process are converging almost-surely to the true parameter. Finally, the consistency is deduced for the common mean, seen as a functional parameter.

Theorem 1.2.6. *Suppose $(\mathbf{Y}_1,\dots,\mathbf{Y}_N)$ is a sequence of iid random variables satisfying the functional nonlinear mixed model observed on fixed time points: for $i = 1,\dots,N$, for $j = 1,\dots,T_i$, $[\mathbf{Y}_i]_j = Y_i(t_{i,j})$. We first make the following assumption, to avoid having to theoretically deal with a modeling bias. We assume that the functions μ, $(U_i)_{i=1,\dots,N}$ and w belong to the space spanned by the considered spline basis, and that $\sigma_\varepsilon \underset{\min T_i \to \infty}{\to} 0$.*

Suppose that the model is identifiable and that the update of $\boldsymbol{\theta}_i$ is a contraction mapping. We assume the existence and positive definiteness of $\mathcal{I}$, which is the limit of minus the expected Hessian matrix of the log-likelihood function based on the model. We also assume that for all $i = 1,\dots,N$, the $r \times r$-matrix $\mathbf{a}_i(\boldsymbol{\theta}_i)$ is non-singular.

Then,

$$((\hat{\boldsymbol{\alpha}}^\mu)^{(\infty)}, (\hat{\sigma}_\varepsilon)^{(\infty)}, (\hat{\underset{\sim}{\Sigma}}^U)^{(\infty)}, \hat{\boldsymbol{\theta}}_0^{(\infty)}, (\hat{\Sigma}^{\boldsymbol{\theta}})^{(\infty)}) \underset{\substack{n\to\infty \\ \min T_i \to \infty}}{\xrightarrow{a.s.}} (\boldsymbol{\alpha}^\mu, \sigma_\varepsilon^2, \underset{\sim}{\Sigma}^U, \boldsymbol{\theta}_0, \Sigma^{\boldsymbol{\theta}}).$$

As a consequence, we get that, from a functional viewpoint, for $\mu \in span(B^\mu)$, if we denote $\hat{\mu} = (\hat{\boldsymbol{\alpha}}^\mu)^{(\infty)} B^\mu$,

$$\|\mu - \hat{\mu}\|_{L_2([0,1])} \underset{\substack{n\to\infty \\ \min T_i \to \infty}}{\xrightarrow{a.s.}} 0.$$

1.2.2 Mixture of segmentation

In this section, we propose to split the considered heterogeneous data into homogeneous clusters of individual curves, each of them being segmented over time into homogeneous regimes. To this end, we consider a mixture of segmentation over the projection of the curves onto some functional basis.

The model and its identifiability

We observe multivariate individual curves $(Y_{ih}(t_j))_{1\le i\le N, 1\le j\le d, 1\le h\le H}$ of dimension H over d timepoints and within a population of size N. The heterogeneous population is studied through a mixture model of K clusters, encoded indifferently in its binary form, $z_{ik} = 1$ if and only if the curve i belongs to the cluster k, and its vector form, $z_i = k$ if and only if the curve i belongs to the cluster k, for $1 \le k \le K$ and $1 \le i \le N$. Each observation belongs to the cluster $k \in \{1,\dots,K\}$ with probability $\pi_k \in [0,1]$. The heterogeneity in time is represented through $L_k + 1$ segments $(I_{k\ell})_{0\le\ell\le L_K}$: if $z_{ik} = 1$ and $j \in I_{k\ell}$, encoded by $w_{j\ell} = 1$,

$$Y_{ih}(t_j) = f_{k\ell h}(t_j) + \eta_{ijh},$$

with η_{ijh} corresponds to some random noise. We propose to decompose our signal into several time periods that are meaningful in practice (in hours, in days, in weeks depending on the application), and to have the same function $f_{k\ell h}$ within the considered interval, through the same segment. The modeling assumption is equivalent to a main function $f_{k\ell h}$ for the hth component, for individuals belonging to the cluster k, and for a timepoint in the ℓth segment. This means that within a segment and a cluster, there is a random variation (seen as a noise) independent and identically distributed over each component of the multivariate curve.

We denote $(\boldsymbol{\alpha}_{ij}) \in \mathbb{R}^p$ the coefficient decomposition vectors of the component $j \in \{1, \ldots, d\}$ onto the functional basis, and the individual $i \in \{1, \ldots, N\}$, and the orthonormal characterization leads to, for the level M,

$$(\mathbf{Y}_{i.}(t_j))_{1 \leq j \leq d} = \Pi \boldsymbol{\alpha}_{ij};$$

where Π is a matrix defined by the functional basis of size M. We consider the wavelet coefficient dataset $(\mathbf{A}_i)_{1 \leq i \leq N} = (\boldsymbol{\alpha}_{i.})_{1 \leq i \leq N} \in (\mathbb{R}^{d \times D})^N$, which defines N observations whose probability distribution is modeled by the following finite matrix-variate Gaussian mixture of segmentation model. For the cluster $k \in \{1, \ldots, K\}$, the heterogeneity in time is described by $L_k + 1$ segments, defined by L_k break-points $T_{k0} < T_{k1} < \cdots < T_{kL_k} < T_{k,L_k+1}$. Then, for an observation $i \in \{1, \ldots, N\}$ in the cluster $k \in \{1, \ldots, K\}$, for $\ell \in \{0, \ldots, L_k\}$ such that $j \in \{T_{k,\ell} + 1, T_{k,\ell} + 2, \ldots, T_{k,\ell+1}\}$, we have:

$$[\mathbf{A}_i]_{j.} | (z_{ik} = 1, W_{j\ell} = 1) = \boldsymbol{\mu}_{k\ell} + \varepsilon_{ij} \tag{1.8}$$

with $\varepsilon_{ij} \sim \mathcal{N}_D(0, \Sigma_{k\ell})$ where $\Sigma_{k\ell}$ is diagonal with the values $(\sigma_{k\ell r})_{1 \leq r \leq D}$.

We first establish the identifiability of the multivariate model (1.8).

Theorem 1.2.7 (Identifiability of (1.8))**.** *Assume that:*

(ID.1) For every $k \in \{1, \ldots, K\}$ and $\ell \in \{0, \ldots, L_k\}$, there exists at least one $r \in \{1, \ldots, p\}$ such that $\sigma_{k\ell r} \neq \sigma_{k,\ell+1,r}$ or $\mu_{k\ell r} \neq \mu_{k,\ell+1,r}$.

(ID.2) We have $D \geq \max_{k \in \{1,\ldots,K\}} L_k + 1$.

(ID.3) If there exists $k \neq k'$ such that $L_k = L_{k'}$ then:

- *there exists $\ell \in \{0, \ldots, L_k\}$ such that $T_{k\ell} \neq T_{k',\ell}$,*
- *or there exists $\ell \in \{0, \ldots, L_k\}$ and $r \in \{1, \ldots, p\}$ such that: $\sigma_{k\ell r} \neq \sigma_{k',\ell,r}$ or $\mu_{k\ell r} \neq \mu_{k',\ell,r}$.*

(ID.4) For every $k \in \{1, \ldots, K\}$, $\pi_k > 0$.

Under these assumptions, the model (1.8) *is identifiable.*

Mixture models are known to be identifiable up to a label switching: two partitions can be the same while the cluster labels being reversed. In this model, a natural order is to choose the labeling of each cluster such that

$$k \leq k' \Leftrightarrow L_k \leq L_{k'}.$$

This alleviates the problem of label switching; and it can be completely removed when the $(L_k)_{1 \leq k \leq K}$ are all different.

Estimation

In this section, we assume that K the number of clusters is known, as well as the number of segment within each cluster $(L_k)_{1 \leq k \leq K}$. Using the model (1.8), under identifiability, by noting $\boldsymbol{T}$ the set of the break points and $\boldsymbol{\theta} = ((\mu_{k\ell r}, \sigma_{k\ell r})_{1 \leq k \leq K, 0 \leq \ell \leq L_K, 1 \leq r \leq p}, (\pi_k)_{1 \leq k \leq K})$ the set of parameters, we obtain the following likelihood:

$$\text{lik}(\mathbf{A}; K, \boldsymbol{T}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \prod_{\ell=0}^{L_k} \prod_{j=T_{k\ell}+1}^{T_{k,\ell+1}} \prod_{r=1}^{D} \left[\frac{1}{\sqrt{2\pi\sigma_{k\ell r}}} e^{-\frac{1}{2\sigma_{k\ell r}} \left(\alpha_{ijr} - \mu_{k\ell r}\right)^2} \right].$$

The mixture model leads to the product over individuals $i \in \{1, \ldots, N\}$ and the sum over the clusters $k \in \{1, \ldots, K\}$ while the segmentation is related to the product over each segment $\ell \in$

$\{0, \dots, L_k\}$ and timepoints indexed by $j \in \{T_{k\ell} + 1, \dots, T_{k,\ell+1}\}$, for the cluster $k \in \{1, \dots, K\}$. In addition to the parameters K, $\boldsymbol{T}$ and $\boldsymbol{\theta}$, we search to estimate the partition $\boldsymbol{z}$. We denote $\hat{\boldsymbol{\theta}}$ the maximum likelihood estimator. We use the *Expectation Maximisation* (EM) algorithm (Dempster et al., 1977) to estimate it.

For $i \in \{1, \dots, n\}, k \in \{1, \dots, K\}$, the computation of $\pi_k^{(c)}$ at step c is explicit. The other parameters $(\mu_{k\ell r}^{(c)}, \sigma_{k\ell r}^{(c)})_{1 \leq k \leq K, 1 \leq \ell \leq L_K, 1 \leq r \leq p}$ are given using the dynamic programming (Bellman and Kalaba, 1957; Kay, 1993).

We then prove that the maximum likelihood estimator is consistent. To simplify notations, we consider univariate functional data or projection of the observed functions onto a 1-dimensional basis, such that $p = 1$ in this section, but the conclusion would be the same. To simplify the notations, we set $\sigma_{k\ell} = 1$, but the results can be extended as well to any variance.

Theorem 1.2.8. *Let* $\mathbf{A}$ *be a matrix of a* $n \times T$ *observations of the model* (1.8) *with true parameter* $\boldsymbol{\theta}^\star, \boldsymbol{T}^\star$ *where the number of clusters* K *and the number of segments* $(L_k)_{1 \leq k \leq K}$ *are known. We assume that there exists* $M > 0$ *such that for all* $k \in \{1, \dots, K\}$ *and* $\ell \in \{0, \dots, L_k\}$,

$$\mu_{k\ell} \in [-M; M];$$

that there exists $\tau_{\min} > 0$ *such that for all* $k \in \{1, \dots, K\}$ *and* $\ell \in \{0, \dots, L_k\}$,

$$T_{k,\ell+1} - T_{k\ell} > \tau_{\min} d.$$

We also assume that $\log(N)/d \underset{n,d \to +\infty}{\longrightarrow} 0$. *If there exists* $k \neq k'$ *such that* $L_k = L_{k'}$ *then we assume that there exists at least* $\tau_{\min} d$ *coordinates* j *such that the distribution of* $Y_{ij} | z_{ik}^\star = 1$ *is different from the distribution of* $Y_{ij} | z_{ik'}^\star = 1$. *Finally, we assume that there exists a constant* $c > 0$ *such that for every* $k \in \{1, \dots, K\}$, $\pi_k > c$, *and Assumption (ID.1). Then,*

$$\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{T}}\right) \underset{n,d \to +\infty}{\overset{\mathbb{P}}{\to}} (\boldsymbol{\theta}^\star, \boldsymbol{T}^\star).$$

Experimental study is given in the main paper, illustrating the good behavior of the method.

1.2.3 Simultaneous confidence bands

While confidence intervals for finite quantities are well-established, constructing confidence bands for objects of infinite dimension, such as functions, poses challenges. In this section, we explore the concept of parametric confidence bands for functional data with an orthonormal basis. Specifically, we revisit the method proposed by Sun and Loader (1994), which yields confidence bands for the projection of the regression function in a fixed-dimensional space. Our contributions are as follows:

- we disentangle the bias issue by explicitly defining the parameter of interest within the approach of Sun and Loader (1994); and
- we propose a method for selecting the dimension of the approximation space, treating it as a model selection problem, with a trade-off between conservatism and confidence level assurance.

Functional regression model

We consider the linear functional model introduced in Equation (1.6). When considering the estimation of the regression function f^{L^*} on the space of dimension L defined by the family $\mathbf{B}^L$, we do not directly estimate f^{L^*} but its projection on this finite space, which corresponds to the projected function $\underline{f}^{L,L^*}(t)$ and its associated coefficients $(\underline{\mu}_\ell^{L,L^*})_{1 \leq \ell \leq L}$.

Proposition 1.2.9. $B^L() \mathbf{P}^L y_i$ *is a Gaussian process with mean* $\underline{f}^{L,L^*}()$ *and covariance function* $(s,t) \mapsto \sigma^2 B^L(s) \mathbf{P}^L \mathbf{B}^{L^\varepsilon} (\mathbf{B}^{L^\varepsilon})^T (\mathbf{P}^L)^T (B^L(t))^T$, *and* $(\underline{\hat{f}}^{L,L^*} - \underline{f}^{L,L^*})()$ *is a centered Gaussian process with covariance function* $C^{L,L^*} : (s,t) \mapsto \frac{\sigma^2}{N} B^L(s) \mathbf{P}^L \mathbf{B}^{L^\varepsilon} (\mathbf{B}^{L^\varepsilon})^T (\mathbf{P}^L)^T B^L(t)^T$.

Even if the estimator $\hat{\underline{f}}^{L,L^*}$ is defined on the functional space associated to $\mathbf{S}^L$, it can also be seen as an estimator of the function f^{L^*} which lies in the space $\mathcal{S}^{L^*}$. In that case, the error includes a functional approximation term due to the approximation of f^{L^*} on the space $\mathcal{S}^L$, which will be nonzero if $L \neq L^*$. It corresponds to the bias of the estimator $\hat{\underline{f}}^{L,L^*}$, i.e. the difference between its expectation and the true f^{L^*}. Indeed, recalling that $f^{L^*} = \underline{f}^{L^*,L^*}$, the error of estimation can be decomposed into

$$\hat{\underline{f}}^{L,L^*}(t) - f^{L^*}(t) = \hat{\underline{f}}^{L,L^*}(t) - \underline{f}^{L,L^*}(t) + \underline{f}^{L,L^*}(t) - \underline{f}^{L^*,L^*}(t) =: Stat_{L,L^*}(t) + Bias_{L,L^*}(t).$$

The first term $Stat_{L,L^*}(t) = \hat{\underline{f}}^{L,L^*}(t) - \underline{f}^{L,L^*}(t)$ is the (unrescaled) statistics of the model. It is a random functional quantity which depends on the estimator $\hat{\underline{f}}^{L,L^*}$. From Proposition 1.2.9, for any $t \in [0,1]$, we define the centered and rescaled statistics $Z_L(t)$ such that:

$$Z_L(t) := \frac{Stat_{L,L^*}(t)}{\sqrt{\mathrm{Var}(Stat_{L,L^*}(t))}} = \frac{\hat{\underline{f}}^{L,L^*}(t) - \underline{f}^{L,L^*}(t)}{\sqrt{C^{L,L^*}(t,t)}} \sim \mathcal{N}(0,1).$$

The covariance function can be estimated using the observations $y_{i.}$ as

$$\hat{C}^{L,L^*}(s,t) = \frac{1}{N-1}\sum_{i=1}^{N}(B^L(s)\mathbf{P}^L y_{i.} - \hat{\underline{f}}^{L,L^*}(s))(B^L(t)\mathbf{P}^L y_{i.} - \hat{\underline{f}}^{L,L^*}(t)).$$

The second term $Bias_{L,L^*}(t) = \mathbb{E}(\hat{\underline{f}}^{L,L^*}(t)) - \underline{f}^{L^*,L^*}(t)$ is the bias of the estimator $\hat{\underline{f}}^{L,L^*}(t)$ when estimating the true function $\underline{f}^{L^*,L^*}(t)$. The bias is due to the fact that the estimation is potentially performed in a different (finite) space than the space where the true function $\underline{f}^{L^*,L^*}$ lives. This is a functional bias, which is not random. It corresponds to the approximation (orthogonal projection if Assumption 1.2.1 holds) of f^{L^*} from $\mathcal{S}^{L^*}$ to the space $\mathcal{S}^L$. It can be written as follows:

$$Bias_{L,L^*}(t) = B^L(t)\underline{\mu}^{L,L^*} - B^{L^*}(t)\mu^{L^*}.$$

Thus, we can deduce that if the family is orthonormal (Assumption 1.2.1 holds), if $L < L^*$, $Bias_{L,L^*}(t) \neq 0$, while if $L \geq L^*$, $Bias_{L,L^*}(t) = 0$.

Confidence Bands of $\underline{f}^{L,L^*}$ for a fixed L The objective is to construct a confidence band for $\underline{f}^{L,L^*}$ based on observations $\mathbf{y}$, for a given $L \in \{L_{\min}, \ldots, L_{\max}\}$. This follows the framework proposed by Sun and Loader (1994), using an unbiased and linear estimator $\hat{\underline{f}}^{L,L^*}$.

Theorem 1.2.10 (Sun and Loader (1994))**.** *Set a probability $\alpha \in [0,1]$. Then, we have*

$$\mathbb{P}\left(\forall t \in [0,1], \left|\hat{\underline{f}}^{L,L^*}(t) - \underline{f}^{L,L^*}(t)\right| \leq \hat{d}^L(t)\right) = 1-\alpha$$
$$\text{with } \hat{d}^L(t) = \hat{q}^L\sqrt{\hat{C}^{L,L^*}(t,t)/N};$$

and $\hat{q}^L$ defined as the solution of the following equation, seen as a function of q^L:

$$\alpha = \mathbb{P}(|t_{N-1}| > q^L) + \frac{\|\tau^L\|_1}{\pi}\left(1 + \frac{(q^L)^2}{N-1}\right)^{-(N-1)/2},$$

with $(\tau^L)^2(t) = \partial_{12}c(t,t) = Var(Z_L(t))'$ where we denote $\partial_{12}c(t,t)$ the partial derivatives of a function $c(t,s)$ in the first and second coordinates and then evaluated at $t = s$.

We can thus deduce a confidence band of level $1-\alpha$ for $\underline{f}^{L,L^*}$:

$$CB_1(\underline{f}^{L,L^*}) = \{\forall t \in [0,1], [\hat{\underline{f}}^{L,L^*}(t) - \hat{d}^L(t); \hat{\underline{f}}^{L,L^*}(t) + \hat{d}^L(t)]\}.$$

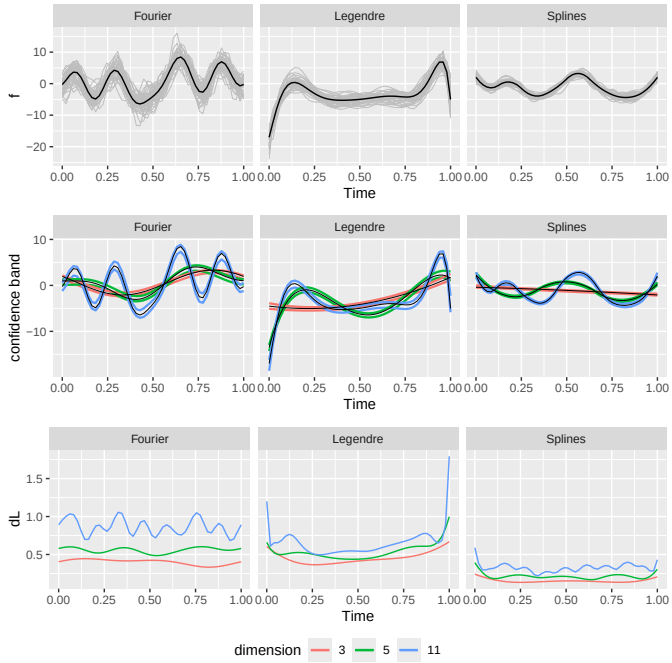


Figure 1.3: Illustrative example. For the three families, resp. Fourier, Legendre and the splines, we display on the top row the observed functional data, on the middle row the confidence bands for different values of L (3, 5 and 11), and on the bottom row the bound dL.

Remark that $\|\hat{d}^L\|_\infty$ increases with L, and when the functions $(B_\ell^L)_{1\leq\ell\leq L}$ consists in an orthonormal family, $\|\hat{d}^L\|_\infty$ increases with L until $L = L^*$ and then $\|\hat{d}^L\|_\infty$ is constant with L. This band is illustrated on Figure 1.3. The top row shows several functional data generated under the Fourier family (left), Legendre (middle), and Spline (right). The middle row displays the confidence bands of $\underline{f}^{L,L^*}$ for different values of $L = 3, 5$ and 11, and the bottom row shows the bound $\hat{d}^L$. The true functions $\underline{f}^{L,L^*}$ are in cyan and the confidence bands are in purple. The bands are very precise for each L, with $\hat{d}^L$ increases with L. As d^L can be seen as a variance, $\hat{d}^L(t)$ is larger on the boundaries of the time domain, as there are fewer observations near 0 and 1.

In the main paper we also evaluate numerically the levels of the obtained confidence bands, which is the expected one whatever the value of L, especially when $L < L^*$ and $L > L^*$ but also when $L > L^\varepsilon$.

Selection of the best confidence band with a criteria taking into account the bias We propose a criterion balancing bias and basis dimension, based on the definition of the band as the estimation of a quantile of an empirical process. Inspired by model selection tools, it helps select the best dimension L. In the following, we assume that $L_{\max}$ is large enough such that $\underline{f}^{L_{\max},L^*} = f^{L^*}$.

We work on the quantile q^L, its oracle version q^{L^*} for the level L^* and the estimation $\hat{q}^L$. All of them are scalar, in a collection of scalars, with $L = L_{\min}, \ldots, L_{\max}$. A natural criteria to choose the best L is such that the estimator $\hat{q}^L$ minimizes the quadratic error $\mathbb{E}\left(\|q^{L^*} - \hat{q}^L\|^2\right)$. However, this quadratic error is unknown as q^{L^*} is unknown. Instead, we study $\|\hat{q}^{L_{\max}} - \hat{q}^L\|^2$. While the theoretical quadratic error $\mathbb{E}\left(\|q^{L^*} - \hat{q}^L\|^2\right)$ decreases when $L < L^*$ and increases when $L > L^*$, the approximation $\|\hat{q}^{L_{\max}} - \hat{q}^L\|^2$ of this error is still decreasing when $L > L^*$.

We observe a bias-like behavior: high when dimension is small and small when dimension is large. Selecting a dimension using this criterion overfits the data. Therefore, we propose penalizing this quantity by the dimension L divided by the sample size N, similar to model se-

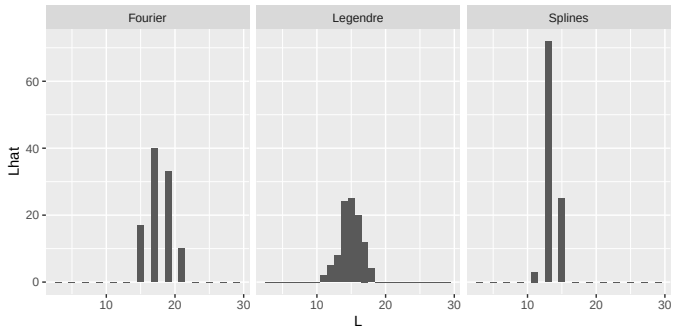


Figure 1.4: Illustrative example. We show the distribution of the selected model, over 100 repetitions, with the new criteria used to select a model for different basis.

lection criteria. We introduce a regularisation parameter $\lambda > 0$ which balances the two terms. A natural criteria to select the best L is then

$$\widetilde{crit}(L) = \|\hat{q}^{L_{\max}} - \hat{q}^{L}\|^2 + \lambda \frac{L}{N}.$$

We define $\tilde{L} = \arg\min_L \widetilde{crit}(L)$, and center the band around $\underline{\hat{f}}^{\tilde{L},L^*}$:

$$CB_2(\underline{f}^{L^*}) = CB_1(\underline{f}^{\tilde{L},L^*})$$

In Figure 1.4, we test which model is selected over 100 repetitions for the three basis. The estimated dimension is equal or larger than the true $L^* = 11$. In the main paper, we have empirically shown that the selected dimension is interesting, and that the related confidence band has a width smaller than the naive confidence band consisting in using $L_{\max}$.

Chapter 2

Network inference by gaussian graphical model and its use

This chapter is the result of collaborations with Mélina Gallopin (I2BC, Institute of Integrative Biology of the Cell), Emeline Perthame (Institut Pasteur), Valérie Monbet (IRMAR, Mathematics Research Institute of Rennes), Madison Giacofci (IRMAR, Mathematics Research Institute of Rennes), Rémi Molinier (Institut Fourier, Mathematics Laboratory) and Marie Morvan (PhD student). Thanks to them!

- Stable network inference in high-dimensional graphical model using single-linkage, Devijver, E., Gallopin, M. and Molinier, R., preprint, 2024+, link HAL.
- Nonlinear network-based quantitative trait prediction from biological data, M. Blein-Nicolas, E. Devijver, M. Gallopin, E. Perthame, Journal of the Royal Statistical Society Series C: Applied Statistics, 2024, link.[a]
- Prediction of the NASH through penalized mixture of logistic regression models, M. Morvan, E. Devijver, M. Giacofci, and V. Monbet, Annals of Applied Statistics, 15(2): 952-970, 2021, link
- Block-diagonal covariance selection for high-dimensional Gaussian graphical models, E. Devijver et M. Gallopin, Journal of the American Statistical Association, 2017, link.[b]

[a] Code available in the R package `xLLiM`
[b] Code available in the R package `shock`

Graphical models (Whittaker, 1990; Lauritzen, 1996) have become a popular tool for representing conditional dependencies among variables using a graph. For Gaussian graphical models (GGMs), the edges of the corresponding graph are the non-zero coefficients of the inverse covariance matrix. To estimate this matrix in high-dimensional contexts, methods based on an ℓ_1-penalized log-likelihood have been proposed (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Banerjee et al., 2008). A popular method is the graphical lasso algorithm introduced by Friedman et al. (2008). In this chapter, we focus first on the graphical lasso algorithm, and then on its use through a more general model. The organization of this chapter is the following:

- In Section 2.1, we focus on the graphical lasso algorithm as an estimator of the Gaussian graphical model in high-dimension. It has been shown that this algorithm can be decomposed into two steps: first, by constructing blocks of independent variables through hierarchical clustering, and then estimating a sparse structure within each block. We propose two theoretical contributions based on this decomposition. The first, detailed in Section 2.1.1, introduces an estimator that focuses on the block-diagonal structure of the graphical lasso by performing model selection on it first and then estimating the sparse structure independently. We provide theoretical results showing that this estimator is adaptively minimax. This is a joint work with Mélina Gallopin, and all the details are available in Devijver and Gallopin (2018). The second contribution, detailed in Section 2.1.2, discusses the stability of the graphical lasso. We derive theoretical bounds that prove that the decomposition into two steps can make the estimator stable. The structure added by the block decomposition enhances stability. This is a joint work with Mélina Gallopin and Rémi Molinier, and all the details are available in Devijver et al. (2024).

- In Section 2.2, we propose two models of prediction, based on graphical modeling, to answer practical questions in biology. This section introduces two methods, the development of which has been driven by data and expert knowledge. The first, detailed in Sectop, 2.2.1, aims to answer the following question: what is the link between a set of ecophysiological traits and the proteomic data of maize? Biomarkers are known to be highly correlated, and the graphical model is needed to account for those correlations. This is a joint work with Mélisande Blein-Nicolas, Mélina Gallopin and Emeline Perthame. All the details are available in Blein-Nicolas et al. (2024). The second, detailed in 2.2.2, aims to answer the following question: can we predict the NASH from blood spectra? The discretization of the spectra is considered raw data, and thus a set of highly correlated covariates. The graphical model is also needed here to account for those correlations. This project corresponds to a part of the PhD thesis of Marie Morvan, with whom I collaborate, and which was supervised by Madison Giacofci and Valérie Monbet. All the details are available in Morvan et al. (2021).

2.1 High-dimensional Gaussian graphical models

Graphical models are a class of statistical models that combine the rigor of probabilistic approaches with the intuitive representation of relationships provided by graphs. They consist of a set of random variables and a graph, where each vertex (or node) represents a random variable and each edge (or link) expresses the dependence structure between the variables. These models are particularly valuable for their interpretability, as the dependence structure among variables is easily readable from the graph.

Considering Gaussian graphical models, the set of variables $\mathbf{Y} = (Y_1, \ldots, Y_p) \in \mathbb{R}^p$, and a sample $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, are drawn from a multivariate normal distribution with density $\phi_p(0, \Sigma)$ where $\Sigma_{j,j} = 1$ for all $j \in \{1, \ldots, p\}$. The edges in the graph are encoded in the precision matrix $\Theta = \Sigma^{-1}$. Thus, to consider a sparse graph, one has to estimate a sparse precision matrix Θ. Several methods have been proposed to penalize the log-likelihood associated to this model (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Banerjee et al., 2008), we focus in this chapter on the Graphical Lasso (Friedman et al., 2008), defined as follows:

$$\hat{\Theta}(\lambda; \mathbf{y}) = \underset{\Theta}{\operatorname{argmax}} \{\log\det\Theta - \operatorname{tr}(S\Theta) - \lambda\|\Theta\|_1\}$$

over nonnegative definite matrices Θ, where λ is a nonnegative tuning parameter and S the sample covariance estimate.

However, these network reconstruction methods often perform poorly in so-called *ultra high-dimensional contexts* (Giraud, 2008; Verzelen, 2012), when the number of observations is much smaller than the number of variables. Even with the ℓ_1 penalty, selecting a relevant level of sparsity is a complex task when the sample size is limited, which is a common situation in various applications, such as in systems biology where the cost of the sequencing technologies may limit the number of available observations (Frazee et al., 2011). In practice, the network reconstruction problem is facilitated by restricting the analysis to a subset of variables, based on external knowledge and prior studies of the data (Ambroise et al., 2009; Yin and Li, 2011). When no external knowledge is available, only the most variable features are typically kept in the analysis (Guo et al., 2011; Allen and Liu, 2013). Choosing the appropriate subset of variables to focus on is a key step in reducing the model dimension and the number of parameters to estimate, but no procedure is clearly established to perform this selection in high-dimensional settings.

Fortunately, Witten et al. (2011) and Mazumder and Hastie (2012) have noticed a particular property of this estimator, that has been very useful for its development: the graphical lasso estimation for a given level of regularization $\lambda \geq 0$ can thus be decomposed into two steps:

Step 1 Identify the connected components of the undirected graph with adjacency matrix A associated to the thresholding of the absolute value of the sample covariance matrix at level λ;

Step 2 Perform Graphical Lasso with parameter λ on each connected component separately.

This decomposition has fasten the algorithm Witten et al. (2011), but also opens a new insight on the graphical lasso estimator. Tan et al. (2015) have noticed that the first step is equivalent to performing a single linkage clustering on the variables. Then, they have proposed the *cluster graphical lasso*, using an alternative to single linkage clustering in the two-step procedure, namely the average linkage clustering. The selection of the cutoff applied to hierarchical clustering in the first step of the *cluster graphical lasso* algorithm is performed independently from the selection of the regularization parameters in the second step of the algorithm. Their results suggest that the detection of the block-diagonal structure of the covariance matrix prior to network inference in each cluster can improve network inference. Other authors have recently proposed procedures to detect the block-diagonal structure of a covariance matrix. Pavlenko et al. (2012) provided a method to detect this structure for high-dimensional supervised classification that is supported by asymptotic guarantees. Hyodo et al. (2015) proposed tests to perform this detection and derived consistency for their method when the number of variables and the sample size tend to infinity. We propose in Section 2.1.1 to come back to the graphical lasso estimation into 2 steps, and to select a model among the model collection from the step 1

before inferring a sparse network in each connected component. Particularly, we prove that the corresponding estimator is adaptive minimax.

Network inference is a domain within statistics where stability is particularly critical. When performing network inference on two data sets derived from the same model with a small sample size using classical methods, the resulting inferred networks are often markedly different. This variability arises from the large number of parameters that need to be estimated and the complexity of the optimization task involved. However, without stability, interpretability becomes challenging, which undermines one of the major advantages of graphical models. This is especially pertinent in the context of regulatory networks derived from real omics data, where observations are typically limited (Frazee et al., 2011; Krumsiek et al., 2011; Michailidis and d'Alché Buc, 2013). As a result, practitioners have often criticized the developed methods, opting instead to manually select an appropriate subset of variables to focus on. However, such external knowledge is not always available and could be enhanced by a deeper understanding of the data and the application of machine learning tools. Several methods have been proposed to stabilize variable selection in GGMs, primarily based on resampling. In Bach (2008); Meinshausen and Buhlmann (2006), the authors suggest subsampling the observations, running a model on each sample, and retaining variables selected consistently across all or most samples. Both papers provide theoretical results that guarantee good performance asymptotically with increasing sample sizes. Building on Bach (2008), Colby et al. (2018) evaluate the stability and accuracy of gene regulatory network inference using bootstrap aggregation. Additionally, Haury et al. (2012), drawing from Bach (2008); Meinshausen and Buhlmann (2006), focuses specifically on bootstrap sampling for network inference. More recently, Bodinier et al. (2023) proposed a score to measure the overall stability of the set of selected features, introducing a new calibration strategy for stability selection. In a broader context, Lim and Yu (2016) introduced `ESCV`, while Bar-Hen and Poggi (2016) proposed removing the most influential observations to achieve stable networks, akin to the jackknife method. However, these methods require substantial computation because they rely on subsampling. Furthermore, large sample sizes are necessary to ensure good performance with subsampling techniques. Then, in Section 2.1.2, we prove that the first step of the graphical Lasso estimator is stable, and illustrate numerically the stability of several methods.

2.1.1 Block-diagonal covariance selection

We propose in this section to recast the estimation of the graphical Lasso into a model selection problem, and propose a non-asymptotic model selection procedure supported by an oracle type inequality and a minimax lower bound, based on the slope heuristic.

Our goal is to infer the graph of conditional dependencies between variables, encoded by the precision matrix $\Theta = \Sigma^{-1}$. Since the matrices Σ and Θ have the same block-diagonal structure, we first seek to detect the optimal block-diagonal structure of the covariance matrix Σ, *i.e.* the optimal partition of variables into blocks. We index the variables from 1 to p. We note $\mathbf{B} = \{\mathbf{B}_1; \ldots; \mathbf{B}_K\}$ the partition of variables into K blocks where $\mathbf{B}_k$ is the subset of variables in block k, and p_k is the number of variables in block k. The partition describes the block-diagonal structure of the matrix: off the block, all coefficients of the matrix are zeros. We consider the following set of multivariate normal densities with block-diagonal covariance matrices:

$$F_{\mathbf{B}} = \left\{ f_{\mathbf{B}} = \phi_p(0, \Sigma_{\mathbf{B}}) \text{ with } \Sigma_{\mathbf{B}} \in \mathbb{S}_p^{++}(\mathbb{R}) \;\middle|\; \begin{array}{l} \lambda_m \leq \min(\mathrm{sp}(\Sigma_{\mathbf{B}})) \leq \max(\mathrm{sp}(\Sigma_{\mathbf{B}})) \leq \lambda_M, \\ \Sigma_{\mathbf{B}} = P_\sigma \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_K \end{pmatrix} P_\sigma^{-1}, \\ \Sigma_k \in \mathbb{S}_{p_k}^{++}(\mathbb{R}) \text{ for } k \in \{1, \ldots, K\} \end{array} \right\}, \tag{2.1}$$

where $\mathbb{S}_p^{++}(\mathbb{R})$ is the set of positive semidefinite matrices of size p, λ_m and λ_M are real numbers, $\min(\mathrm{sp}(\Sigma_{\mathbf{B}}))$, $\max(\mathrm{sp}(\Sigma_{\mathbf{B}}))$ are the smallest and highest eigenvalues of $\Sigma_{\mathbf{B}}$ and P_σ is a permutation matrix leading to a block-diagonal covariance matrix. We denote $D_{\mathbf{B}} = \sum_{k=1}^{K} p_k(p_k - 1)/2$ the dimension of the model $F_{\mathbf{B}}$.

As the set of all possible partitions of variables is large (its size is the Bell number), we

consider the collection:

$$\mathcal{B}_\Lambda = (\mathbf{B}_\lambda)_{\lambda\in\Lambda} \tag{2.2}$$

given by the partition of variables corresponding to the block-diagonal structure of the adjacency matrix $E_\lambda = [\mathbf{1}_{\{|S_{j,j'}|>\lambda\}}]_{\substack{1\le j\le p\\ 1\le j'\le p}}$, based on the thresholded absolute value of the sample covariance matrix S. Note that the data is scaled if needed so that the set of thresholds $\Lambda \subset [0,1]$ covers all possible partitions derived from E_λ.

Once we have constructed the collection of models $\mathcal{F}_\Lambda = (F_\mathbf{B})_{\mathbf{B}\in\mathcal{B}_\Lambda}$, we select a model among this collection using the following model selection criterion:

$$\hat{\mathbf{B}} = \underset{\mathbf{B}\in\mathcal{B}_\Lambda}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(\hat{f}_\mathbf{B}(\mathbf{y}_i)) + \operatorname{pen}(\mathbf{B}) \right\},$$

where $\operatorname{pen}(\mathbf{B})$ is a penalty term to define and $\hat{f}_\mathbf{B} = \phi_p(0, \hat{\Sigma}_\mathbf{B})$ where $\hat{\Sigma}_\mathbf{B}$ is the maximum likelihood estimator of $\Sigma_\mathbf{B}$. The matrix $\hat{\Sigma}_\mathbf{B}$ is constructed block by block, using the sample covariance matrix of the dataset restricted to variables in each block.

The penalty term $\operatorname{pen}(\mathbf{B})$ is based on non-asymptotic model selection properties. We aim at selecting, among $\mathcal{B}$, the optimal partition $\mathbf{B}^\star$. First, for each model indexed by $\mathbf{B}$, we consider the density $\hat{f}_\mathbf{B} = \phi_p(0, \hat{\Sigma}_\mathbf{B})$ where $\hat{\Sigma}_\mathbf{B}$ is the maximum likelihood estimator of $\Sigma_\mathbf{B}$. Among all $\mathbf{B} \in \mathcal{B}_\Lambda$, we want to select the density $\hat{f}_\mathbf{B}$ which is the closest one to the true distribution $f^\star$. To measure the distance between the two densities, we define the risk:

$$R_\mathbf{B}(f^\star) = \mathbb{E}(d^2(f^\star, \hat{f}^\mathbf{B})),$$

where d is a distance between two densities. Ideally, we would like to select the partition $\mathbf{B}$ that minimizes the risk $R_\mathbf{B}(f^\star)$: this partition is called the oracle. Unfortunately, it is not reachable in practice because the true density $f^\star$ is unknown. However, we will prove that we do almost as well as the oracle, i.e. we select a model for which the risk of the procedure is upper bounded by the oracle risk, up to a constant.

Before stating the theorem, we recall the definition of the Hellinger distance between two densities f and g defined on $\mathbb{R}^p$, $\mathrm{d}_H^2(f,g) = \frac{1}{2}\int_{\mathbb{R}^p}(\sqrt{f(x)} - \sqrt{g(x)})^2 dx$, and the Kullback-Leibler divergence between two densities f and g defined on $\mathbb{R}^p$, $\mathrm{KL}(f,g) = \int_{\mathbb{R}^p} \log\left(\frac{f(x)}{g(x)}\right) f(x)dx$.

Theorem 2.1.1. *Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ be the observations, arising from a density $f^\star$. Let $\mathcal{B}_\Lambda \subset \mathcal{B}$ as defined in* (2.2)*., and the collection of models $\mathcal{F}_\Lambda = (F_\mathbf{B})_{\mathbf{B}\in\mathcal{B}_\Lambda}$. We denote by $\hat{f}_\mathbf{B}$ the maximum likelihood estimator for the model $F_\mathbf{B}$. Let $\tau > 0$, and for all $\mathbf{B} \in \mathcal{B}$, let $f_\mathbf{B} \in F_\mathbf{B}$ such that:*

$$\begin{aligned} \mathrm{KL}(f^\star, f_\mathbf{B}) &\le 2 \inf_{f\in F_\mathbf{B}} \mathrm{KL}(f^\star, f); \\ f_\mathbf{B} &\ge \exp(-\tau) f^\star. \end{aligned} \tag{2.3}$$

Then, there exists some absolute constants κ and C_{oracle} such that whenever

$$\operatorname{pen}(\mathbf{B}) \ge \kappa \frac{D_\mathbf{B}}{n}\left[2c^2 + \log\left(\frac{p^4}{D_\mathbf{B}(\frac{D_\mathbf{B}}{n}c^2 \wedge 1)}\right)\right]$$

for every $\mathbf{B} \in \mathcal{B}$, with $c = \sqrt{\pi} + \sqrt{\log(3\sqrt{3}\frac{\lambda_M}{\lambda_m})}$, the random variable $\hat{\mathbf{B}} \in \mathcal{B}_\Lambda$ such that

$$\hat{\mathbf{B}} = \underset{\mathbf{B}\in\mathcal{B}_\Lambda}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(\hat{f}_\mathbf{B}(\mathbf{y}_i)) + \operatorname{pen}(\mathbf{B}) \right\}$$

exists and, moreover, whatever the true density $f^\star$,

$$\mathbb{E}(\mathrm{d}_H^2(f^\star, \hat{f}_{\hat{\mathbf{B}}})) \le C_{oracle}\mathbb{E}\left[\inf_{\mathbf{B}\in\mathcal{B}_\Lambda}\left(\inf_{f\in F_\mathbf{B}} \mathrm{KL}(f^\star, f) + \operatorname{pen}(\mathbf{B})\right)\right] + \frac{1\vee\tau}{n} p\log(p). \tag{2.4}$$

This theorem is deduced from an adaptation of a general model selection theorem for maximum likelihood estimator developed by Massart (2007). To use this theorem, the main assumptions to satisfy are the control of the bracketing entropy of each model in the whole collection of models and the construction of weights for each model to control the complexity of the collection of models. To compute the weights, we use combinatorics arguments. The control of the bracketing entropy is a classical tool to bound the Hellinger risk of the maximum likelihood estimator, and has already been done for Gaussian densities in Genovese and Wasserman (2000) and Maugis and Michel (2011).

The assumption on the true density $f^\star$ (2.3) is done because we consider a random subcollection of models $\mathcal{F}_\Lambda$ from the whole collection of models $\mathcal{F}$. Thanks to this assumption, we use the Bernstein inequality to control the additional randomness. The parameter τ depends on the true unknown density $f^\star$ and cannot be explicitly determined for this reason. We could do some hypothesis on the true density $f^\star$ to be able to explicit τ but we choose not to do it: e.g. under the assumption that the Kullback-Leibler divergence and the Hellinger distance are equivalent, we can explicitly determine τ. Note that the parameter τ only appears in the rest term $\mathbf{r} = \frac{1\vee\tau}{n} p\log(p)$ and not on the penalty term $\mathrm{pen}(\mathbf{B})$. Therefore, we do not need to explicit τ to select a model.

We remark that the Hellinger risk is upper bounded by the Kullback-Leibler divergence in (2.4). For this reason, the result (2.4) is not exactly an oracle inequality and is called an oracle type inequality. However, the use of the Kullback-Leibler divergence and the Hellinger distance is common for model selection theorem for Maximum Likelihood Estimator: e.g. Theorem 7.11 in Massart (2007). Moreover, the Kullback-Leibler divergence is comparable to the Hellinger distance under some assumptions. Under these assumptions, the result (2.4) is exactly an oracle inequality.

The collection of models (2.1) is defined such that covariance matrices have bounded eigenvalues. These bounds are useful to control the complexity of each model by constructing a discretization of this space. Every constant involved in (2.4) depends on these bounds. This assumption is common in non-asymptotic model selection framework.

To complete this analysis, we provide a second theoretical guarantee. In contrast with Lebarbier (2005); Maugis and Michel (2011), we strengthen the oracle type inequality using a minimax lower bound for the risk between the true model and the model selected. Note that in Gaussian Graphical Models, lower bounds have already been obtained in other contexts (Bickel and Levina, 2008; Cai et al., 2010).

In Theorem 2.1.1, we have proved that we select a model as good as the oracle model in a density estimation framework. However, the bound has two extra terms: the penalty term $\mathrm{pen}(\mathbf{B})$ and the rest, which give the rate of the estimator. Based on Theorem 2.1.1 only, we do not know if the rate is as good as possible. The following theorem lower bounds the risk by a rate with the same form as the upper bound (seen as a function of n, p and $D_{\mathbf{B}}$), which guarantees that we obtain an optimal rate.

Theorem 2.1.2. *Let $\mathbf{B} \in \mathcal{B}$. Consider the model $F_{\mathbf{B}}$ defined in* (2.1)*, and $D_{\mathbf{B}}$ its dimension. Then, if we denote $C_{minim} = \frac{e}{4(2e+1)^2(8+\log(\lambda_M/\lambda_m))}$, for any estimator $\hat{f}_{\mathbf{B}}$ of $f^\star$ one has*

$$\sup_{f^\star \in F_{\mathbf{B}}} \mathbb{E}(\mathrm{d}_H^2(\hat{f}_{\mathbf{B}}, f^\star)) \geq C_{minim} \frac{D_{\mathbf{B}}}{n} \left(1 + \log\left(\frac{2\lambda_M p(p-1)}{D_{\mathbf{B}}}\right)\right).$$

To obtain this lower bound, we use Birgé's lemma (Birgé, 2005) in conjunction with a discretization of each model, already constructed to obtain the oracle type inequality. We assume that the parameters of the models in the collection (2.1) are bounded, which is not a strong assumption. The constant involved is explicit.

Thanks to Theorems 2.1.1 and 2.1.2, we upper bound and lower bound the Hellinger risk, proving that our procedure is adaptive minimax. Note that the model selection procedure is optimized for density estimation and not for edge selection: the Hellinger distance and Kullback-Leibler divergence measure the differences between two densities from an estimation point of view. In contrast, network inference focuses on edge selection. However, we point out that the model selection procedure is only proposed in a specific context (n small), as a preliminary step prior to edge selection . Although this preliminary step is not optimized for

edge selection, it improves the network inference procedure as illustrated in simulated data. To conclude, let recall that these results are non-asymptotic, which means that they hold for a fixed sample size n, which is particularly relevant in a context where the number of observations n is limited. The results are consistent with the point of view adopted in this work.

In practice, we consider a simpler version of the penalty term:

$$\text{pen}(\mathbf{B}) = \kappa \frac{D_{\mathbf{B}}}{n} \tag{2.5}$$

where κ is a constant depending on absolute constants and on the bounds λ_m and λ_M. Such simplification has already been proposed by Lebarbier (2005). The extra term $\frac{D_{\mathbf{B}}}{n} \log\left(\frac{p(p-1)}{D_{\mathbf{B}}}\right)$ is useful to overpenalize the collection of models when it contains many models with the same sizes. The simplification of the penalty term (2.5) is reasonable for moderate number of variables.

Subsequently, we note that the bounds λ_m and λ_M are non-tractable. For this reason, we prefer to calibrate the constant κ in (2.5) from the data. This calibration is based on the slope heuristic, originally proposed and proved in the context of heteroscedastic regression with fixed design (Birgé and Massart, 2007; Baraud et al., 2009), and for homoscedastic regression with fixed design (Arlot and Massart, 2009). In other contexts, the slope heuristic has been used and have proven to be effective for multiple change point detection (Lebarbier, 2005), for variable selection in mixture models (Maugis and Michel, 2011), for choosing the number of components in Poisson mixture models (Rau et al., 2015) or for selecting the number of components in discriminative functional mixture models (Bouveyron et al., 2015).

Baudry et al. (2012) have provided practical tools to calibrate the coefficient κ in (2.5) based on the slope heuristic developed by Birgé and Massart (2007). Note that the detection of the optimal $\mathbf{B}$ is easy to implement in practice and does not rely on heavy computation such as cross-validation techniques.

Once we have detected the optimal block-diagonal structure of the GGM, network inference is performed independently in each block using the graphical lasso (Friedman et al., 2008).

2.1.2 Stable network inference using single-linkage

In this section, we argue that the decomposition into two steps of the Graphical Lasso problem enhances the stability of network inference. We experimentally illustrate this improvement and theoretically prove that single linkage is stable, whereas other classical linkages, such as average linkage, are not.

Theoretical result for the stability of the modular decomposition

In this section, we are interested in the stability of the hierarchical clustering, in the sense that, if two samples are observed generated from the same distribution, we want to measure how close are the two dendograms provided by the hierarchical clustering. Let $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ and $(\mathbf{y}_1, \ldots, \mathbf{y}_{i-1}, \tilde{\mathbf{y}}_i, \mathbf{y}_{i+1} \ldots, \mathbf{y}_n)$ be two samples in $\mathbb{R}^p$ from the same multivariate normal distribution with density $\phi_p(\mathbf{0}, \Sigma)$ where $\Sigma_{j,j} = 1$ for all $j \in \{1, \ldots, p\}$. We assume that observations are standardized, and we focus on empirical correlations matrices.

A dendogram is an right-continuous application $\theta : [0, \infty) \to \mathcal{P}(\{1, \ldots, p\})$ that defines a nested family of partitions of $\{1, \ldots, p\}$ which starts with only singletons and ends with the whole space. We denote Θ_p the set of all dendrograms on $\{1, \ldots, p\}$.

We work here with ultrametrics, that are associated to dendrograms through a one-to-one mapping. An ultrametric space (X, u) is a metric space which satisfies a stronger type of triangle inequality: for all $(x, x', x'') \in X^3$,

$$\max(u(x, x'), u(x', x'')) \geq u(x, x'').$$

For a finite set $\{1, \ldots, p\}$, we denote $\mathcal{U}_p$ the set of all ultrametrics on $\{1, \ldots, p\}$.

Theorem 9 in Carlsson and Mémoli (2010) gives a one to one correspondence

$$\Psi \colon \Theta_p \to \mathcal{U}_p$$

where, for $\theta \in \Theta_p$, $u = \Psi(\theta)$ is the ultrametric on $\{1, \dots, p\}$ defined for all $(x, y) \in \{1, \dots, p\}^2$ by

$$u(x, y) = \min \{t \geq 0 \mid x \text{ and } y \text{ are in the same subset in the partition } \theta(t)\}.$$

Note that $u = \Psi(\theta)$ is also, by definition, the cophenetic distance associated to the dendogram θ: $u(i, j)$ corresponds to the height at which stage i and j are merged together. We compare those cophenetic distances for two dendograms using the following distance.

Definition 2.1.3. *The distance d_{coph} is defined by, for two dendograms $\theta_1, \theta_2 \in \Theta_p$, and their associated ultrametrics $u_1 = \Psi(\theta_1)$ and $u_2 = \Psi(\theta_2)$,*

$$d_{coph}(\theta_1, \theta_2) = \max_{1 \leq i,j \leq p} |u_1(i, j) - u_2(i, j)|.$$

The inverse $\theta = \Psi^{-1}(u)$ for $u \in \mathcal{U}_p$ is given, for $t \geq 0$, by $\theta(t)$ to be the partition obtained from the equivalence relation $\sim_{u,t}$ where, for $(x, y) \in \{1, \dots, p\}^2$,

$$x \sim_{u,t} y \quad \iff \quad u(x, y) \leq t.$$

We will denote by C_p the complete simple graph with $\{1, \dots, p\}$ as set of vertices and a path in C_p with ν vertices will be encoded by a map $\eta \colon \{1, 2, \dots, \nu\} \to \{1, 2, \dots, p\}$ where, for all $k \in \{1, 2, \dots \nu\}$, $\eta(\nu)$ yields the kth vertex of the path. We introduce in the next definition the application u_A and then show that it is an ultrametric on $\{1, \dots, p\}$.

Definition 2.1.4. *Let $A \in \mathbb{R}^{p \times p}$ be a symmetric matrix with positive nondiagonal entries, and zeros on the diagonal. We define the following application.*

$$u_A \colon \{1, \dots, p\}^2 \longrightarrow \mathbb{R}$$
$$(i, j) \longmapsto \begin{cases} 0 \text{ if } i = j, \\ \min\limits_{\eta \text{ a path from } i \text{ to } j \text{ in } C_p} \left\{ \max\limits_k (A_{\eta(k), \eta(k+1)}) \right\} \text{ elsewhere.} \end{cases}$$

If $A \in \mathbb{R}^{p \times p}$ is a symmetric matrix with positive nondiagonal entries, and zeros on the diagonal, the application u_A defines an ultrametric on $\{1, \dots, p\}$, and we denote by $\theta_A = \Psi^{-1}(u_A)$ the dendrogram associated to the ultrametric u_A.

Remark that, if the matrix A is associated to a distance d, the dendogram θ_A is exactly the one obtained by the single linkage hierarchical clustering with the distance d (see Carlsson and Mémoli (2010, Corollary 14)). One can particularly use $A = \mathbf{1} - |S^1|$, with S^1 the sample covariance matrix and $\mathbf{1}$ corresponds to the matrix with 1 for each coefficient, which is the one constructed in the first step of the Graphical Lasso (Tan et al., 2015).

The main theoretical contribution of this section is the following theorem, which gives the stability of the dendrogram constructed in the first step of the Graphical Lasso.

Theorem 2.1.5. *Let two samples $(\mathbf{y}_1, \dots, \mathbf{y}_n)$ and $(\mathbf{y}_1, \dots, \tilde{\mathbf{y}}_i, \dots, \mathbf{y}_n)$ where $\tilde{\mathbf{y}}_i \sim \mathbf{Y}$ and are iid, and S and $\tilde{S}$ the corresponding sample covariance matrices. Then, for $\alpha \in (0, 1)$, with probability $1 - \alpha$,*

$$d_{coph}(|\theta_{\mathbf{1}-|S|}|, \theta_{\mathbf{1}-|\tilde{S}|}) \leq \frac{2p}{(n-1)\sqrt{\alpha}}.$$

Asymptotically, the two collections of modules detected by the Graphical Lasso on two samples where only one observation differs, varying the regularization parameter λ, are the same. This means that the single linkage used in the first step of the Graphical Lasso is a good choice, with respect to the stability of the collection of models that is considered.

Similarly to Carlsson and Mémoli (2010, Remark 17), we can show that the complete linkage and the average linkage are unstable: small perturbations of the matrix A may lead to large perturbations of the corresponding ultrametric.

Table 2.1: Stability of dendrograms (measured with the normalized distance coph_n and its standard deviation in parenthesis) using hierarchical clustering with several linkages and ARI between the clusters get by hierarchical clustering using several linkages and several model selection criterion. Best results are bolded.

		AL	CL	ML	SL	WL
coph_n		0.81 (0.08)	0.91 (0.05)	0.85 (0.06)	**0.59** (0.12)	1.01 (0.10)
ARI	2	0.10	0.02	0.05	0.09	**0.78**
	SH	0.49	0.32	0.44	**0.57**	0.37
	BIC	0.26	0.11	0.17	**0.39**	0.27

Experiments

In the main paper, we illustrated the stability of dendrograms using hierarchical clustering with varying linkages (Theorem 2.1.5), the stability of clusters determined by cutting the dendrogram with a model selection criterion, and the stability of network inference on simulated data and two real datasets. Here, we focus on the BRCA gene expression dataset for breast cancer patients, measured with RNA-Sequencing and generated by the TCGA Research Network. The data, downloaded from TCGA portals using the TCGA2STAT tool Wan et al. (2015), consists of 1212 samples and 9191 genes. We focus on the 200 most variable genes, constructing 17 batches of 70 samples each, resulting in 1190 observations.

Stability of hierarchical clustering: which linkage method? In this section, we validate the theoretical results from Section 2.2.2. We compare the stability of dendrograms generated by different hierarchical clustering linkage methods: average (AL), complete (CL), McQuitty (ML), single (SL), and Ward (WL). The measure used is the distance introduced in Definition 2.1.3, which we normalize to facilitate the analysis: for two matrices A_1, A_2, and their associated dendograms $\theta_{A_1}, \theta_{A_2}$,

$$d^N_{\text{coph}}(\theta_{A_1}, \theta_{A_2}) = \max_{1 \leq i,j \leq p} \left| \frac{u_{A_1}(i,j)}{\max(u_{A_1})} - \frac{u_{A_2}(i,j)}{\max(u_{A_2})} \right|.$$

Table 2.1 shows the stability of dendrograms for generated data, BRCA, and equities, with normalized distance and standard deviation. In the BRCA dataset, SL shows the highest stability, indicating robustness for biological data with high sample variability.

Next, we cut the dendrograms to focus on clustering, treating this as a model selection problem using the Bayesian Information Criterion (BIC, Schwarz (1978)) and the slope heuristic (SH, Birgé and Massart (2001); Baudry et al. (2012)). Table 2.1 presents the Adjusted Rand Index (ARI) for clusters derived from different linkage methods and model selection criteria, where an ARI of 1 indicates a perfect match. SL and WL are the most competitive, with WL being the most stable but not sparse, followed closely by SL combined with SH.

These results show that the choice of linkage method and model selection criterion significantly impacts clustering stability and accuracy, with SL and SH generally providing the most stable results.

Stability of inferred networks In this section, we evaluate the stability of networks inferred by classical methods using the normalized Hamming distance between two graphs G_1 and G_2, with adjacency matrices A_1 and A_2:

$$d_H(G_1, G_2) = \frac{2\|A_1 - A_2\|_1}{\|A_1\|_1 + \|A_2\|_1}.$$

This metric measures the difference between two graphs, normalized by the total number of edges. We also report the density of the inferred graphs and the CPU time required for computations.

We compare the following Graphical Lasso-based strategies:

Table 2.2: Performance on BRCA. We compare the density, the performance in stability (evaluated by the normalized Hamming distance Hamm) and the computation time (evaluated by the CPU time).

1 step	BIC	EBIC	STARS	ESCV	BL	SS
Dens	0.05	0.00	0.09	0.00	0.01	0.00
Hamm	0.19	0.00	0.20	0.00	0.02	0.01
CPU	73	73	1621	971	1071	7919
2 steps	$\text{SL-SH}_{\text{BIC}}$	$\text{SL-SH}_{\text{EBIC}}$	$\text{SL-SH}_{\text{STARS}}$	$\text{SL-SH}_{\text{ESCV}}$	SL-SH_{BL}	$\text{AL-2}_{\text{sparse}}$
Dens	0.02	0.02	0.01	0	0.01	0.26
Hamm	0.04	0.04	0.02	0	0.01	0.68
CPU	118	14	250	560	2372	2

- One-step Graphical Lasso methods, where the regularization parameter is selected by BIC, EBIC (with $\gamma = 0.5$), STARS and ESCV.
- Stabilized methods based on the one-step Graphical Lasso: BoLasso (BL), Stability Selection (SS)
- Two-step Graphical Lasso methods: based on single linkage, cut with the slope heuristic, and with regularization parameter selected by BIC, STARS, ESCV, and BoLasso[1] within each module; or mimicking CGL based on average linkage with 2 clusters, where in each module the sparser model is selected.

Table 2.2 presents the performance in terms of density, normalized Hamming distance (Hamm), and computation time (CPU time in seconds).
For one-step methods, ESCV and EBIC show very high stability (Hamming distance is zero) but infer empty networks (density is zero), limiting their practical utility. STARS outperforms BIC and BoLasso in network estimation but at higher computational costs. BoLasso and SS show promise in stability and network estimation quality but are computationally intensive, especially SS. For two-step methods, single linkage with SH improves stability and estimation performance over one-step methods. CGL is generally unstable, confirming theoretical expectations about average linkage methods. Overall, two-step methods improve network estimation and stability across both datasets, addressing limitations of one-step methods.

In conclusion, while one-step methods like STARS show competitive performance, especially in estimation accuracy, two-step methods, particularly those using single linkage with appropriate selection criteria, offer superior stability and estimation quality at increased computational costs. These findings underscore the importance of method selection based on both performance metrics and computational feasibility in practical applications of graphical model inference.

[1]Stability Selection was not run in the two-step Graphical Lasso methods due to high computational cost.

2.2 Prediction using a network

In many applications, multivariate data have a graphical structure, and when using them to predict a response, it is necessary to model this structure. Nonparametric methods have been well-used because we don't need to specify the model, but in some cases, it is important to get an interpretable model, and thus parametric forms are preferred. Therefore, it is necessary to incorporate the graphical model introduced in Section 2.1 into prediction methods, such as the one introduced in Chapter 1. In this section, we discuss two contributions that are completely driven by the biological application. The analysis of multivariate biological data is a challenging task, and extensive efforts have been made to provide a wide range of methods to extract information from the data. Observing the data, this network structure is inherent and should be modeled. We do not pretend here to propose generic method that can be used for general datasets, but we argue that the model has been well specified to the data, detail in this section how, and provide interpretable tools to engage the discussions with biologists. We do not pretend here to propose generic methods that can be used for general datasets, but we argue that the model has been well specified to the data, and we detail in this section how, and provide interpretable tools to engage in discussions with biologists.

The first contribution is answering a regression task. We aim to predict a quantitative phenotype (namely ecophysiological traits) from biomarkers (genotypes of dent maize), using a dataset published in Blein-Nicolas et al. (2020). It is known that the link function relating the quantitative trait to explanatory variables is potentially complex and therefore nonlinear (Torres-García et al., 2009), while biologists need interpretable models to understand and analyze the prediction. Additionally, in omic data, it is known that biological entities interact through an unknown module-structured regulatory network Barabási et al. (2011), and no method addresses the regression problem while integrating the regulation network between predictors. We propose in Section 2.2.1 a method to predict a multivariate continuous response from a large set of covariates that are correlated through a modular structure.

The second contribution is a classification task. We aim to predict if a patient has NASH (Non-Alcoholic Steatohepatitis) from mid-infrared spectra. Non-Alcoholic Fatty Liver Disease (NAFLD) is a leading cause of liver disease in Western countries, affecting about 24% of the population (Younossi et al., 2018a) and progressing to the more severe NASH. Diagnosing NASH is challenging due to its asymptomatic nature and the need for invasive liver biopsies, with no current consensus on non-invasive diagnostic methods (see *e.g.* Younossi et al., 2018b). Mid-infrared spectroscopy offers a molecular fingerprint of body fluids and holds potential for predicting and understanding disease consequences. However, mid-infrared spectra, representing absorbance of biological samples over various wavelengths, present several statistical challenges: high-dimensional framework due to the functional aspect of spectra, graph structure among the observations of the spectra, and inter-individual variability due to external or metabolic factors related to the pathology. We propose in Section 2.2.2 a method to predict a binary response (having NASH or not) from a large set of covariates that are correlated.

2.2.1 Nonlinear network-based quantitative prediction from biological data

Quantitatively predicting phenotypic variables using biomarkers is challenging for several reasons. Biological observations might be heterogeneous, reflecting different mechanisms, and the biomarkers used for prediction can be highly correlated due to unknown regulatory networks. In this section, we present a novel approach to predict multivariate quantitative traits from biological data, addressing both issues. The proposed model not only performs well in prediction but is also fully parametric, with clusters of individuals and regulatory networks, facilitating downstream biological interpretation.

A network-based prediction model

Consider a multivariate response $\mathbf{Y} \in \mathbb{R}^L$ and a set of covariates $\mathbf{X} \in \mathbb{R}^D$. To capture the nonlinear relationship between the response and covariates and to model heterogeneous populations, we approximate the regression function with a mixture of K affine regression functions considered as several clusters. The latent variable Z describes cluster membership: $Z_i = k$ if

the individual $i \in \{1, \dots, n\}$ originates from cluster $k \in \{1, \dots, K\}$. Then, we define the direct regression with parameter $(\boldsymbol{A}_k^*, \boldsymbol{b}_k^*, \boldsymbol{c}_k^* \boldsymbol{\Sigma}_k^*, \boldsymbol{\Gamma}_k^*)_{1 \leq k \leq K}$ by

$$\begin{aligned} \mathbf{Y}_i &= (\boldsymbol{A}_k^* \mathbf{X}_i + \boldsymbol{b}_k^* + \varepsilon_i) \mathbb{1}_{Z_i = k} \\ \varepsilon_i | Z_i = k &\sim \mathcal{N}_L(0, \boldsymbol{\Sigma}_k^*) \\ \mathbf{X}_i | Z_i = k &\sim \mathcal{N}_D(\boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*). \end{aligned}$$

If D is large, the least square estimator associated to this model is not identifiable, and one can relate to the corresponding inverse regression with parameter $(\boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{c}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Gamma}_k)_{1 \leq k \leq K}$, introduced in Section 1.1.1:

$$\begin{aligned} \mathbf{X}_i &= (\boldsymbol{A}_k \mathbf{Y}_i + \boldsymbol{b}_k + e_i) \mathbb{1}_{Z_i = k} \\ e_i | Z_i = k &\sim \mathcal{N}_D(0, \boldsymbol{\Sigma}_k) \\ \mathbf{Y}_i | Z_i = k &\sim \mathcal{N}_L(\boldsymbol{c}_k, \boldsymbol{\Gamma}_k). \end{aligned}$$

The joint distribution of $(\mathbf{Y}, \mathbf{X})$ is then a finite mixture of multivariate Gaussian distributions. This model is directly inherited from GLLiM, the model developed in Deleforge et al. (2015). Using the inverse regression trick means that we estimate the inverse regression parameters $(\boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{c}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Gamma}_k)_{1 \leq k \leq K}$ and use the parameters $(\boldsymbol{A}_k^*, \boldsymbol{b}_k^*, \boldsymbol{c}_k^*, \boldsymbol{\Sigma}_k^*, \boldsymbol{\Gamma}_k^*)_{1 \leq k \leq K}$ in the regression of interest for prediction.

There are many covariates compared to the potentially low number of individuals in each cluster, requiring the estimation of numerous parameters, especially covariance matrices. Additionally, biomarkers such as proteins or genes interact through unknown regulatory networks linked to the phenotypic response. This implies that, conditionally on the phenotypic response, each variable interacts with only a few others, forming small modules of correlated variables. The matrix $\boldsymbol{\Sigma}_k$ represents the residual covariance of the covariates $\mathbf{X}$ conditionally on $\mathbf{Y}$ for the cluster $k \in \{1, \dots, K\}$, then we assume that $\boldsymbol{\Sigma}_k$ has a block-diagonal structure, up to a permutation. Interestingly, this leads to a decomposition of $\boldsymbol{\Gamma}_k^*$ into a sum of a block-diagonal matrix $\boldsymbol{\Sigma}_k$ and a low rank matrix described by $\boldsymbol{A}_k \boldsymbol{\Gamma}_k^{1/2}$. For a given cluster $k \in \{1, \dots, K\}$, we decompose $\boldsymbol{\Sigma}_k$ into G_k blocks: for the cluster $k \in \{1, \dots, K\}$,

$$\boldsymbol{\Sigma}_k(B_k) = P_k \begin{pmatrix} \boldsymbol{\Sigma}_k^{[1]} & 0 & \dots & 0 \\ 0 & \boldsymbol{\Sigma}_k^{[2]} & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \ddots & \boldsymbol{\Sigma}_k^{[G_k]} \end{pmatrix} P_k^{-1};$$

where $d_k^{[g]}$ is the set of variables into the gth group, for $g \in \{1, \dots, G_k\}$, $\#\{d_k^{[g]}\}$ the number of variables in the corresponding set, $B_k = (d_k^{[1]}, \dots, d_k^{[G_k]})$, P_k corresponds to the permutation matrix in cluster k, and $\boldsymbol{\Sigma}_k^{[g]} \in \mathcal{S}^{++}_{\#\{d_k^{[g]}\}}(\mathbb{R})$ corresponds to the residual correlations between the $\#\{d_k^{[g]}\}$ variables in group $g \in \{1, \dots, G_k\}$, with $\mathcal{S}_d^{++}(\mathbb{R})$ denotes the space of positive definite matrices of size d with real entries. Remark that each set of groups is specific to each cluster of individuals.

Using the inverse regression trick gives the identifiability of the model (up to label switching), both as a mixture of linear regressions with Gaussian distributions and as a sparse residual covariance matrix model in the inverse regression. Notice that if one wants to estimate the model with $(\boldsymbol{\Gamma}_k^\star)_{1 \leq k \leq K}$ decomposed into sparse + low rank, the problem is ill-posed and intractable (see Chandrasekaran et al. (2011); Candès et al. (2009) for some general solution).

The prediction of a new response $\hat{\mathbf{Y}}_{n+1}$ from a new covariate $\hat{\mathbf{x}}_{n+1}$ is computed afterwards by a linear combination of the linear models associated to each cluster such as we have:

$$\hat{\mathbf{Y}}_{n+1} = \mathbb{E}(\mathbf{Y}_{n+1} | \mathbf{X}_{n+1} = \mathbf{x}_{n+1}) = \sum_{k=1}^{K} w_k^*(\mathbf{x}_{n+1}) \left(\boldsymbol{A}_k^* \mathbf{x}_{n+1} + \boldsymbol{b}_k^* \right)$$

$$w_k^*(\mathbf{x}) = \frac{\pi_k^* \varphi_D(\mathbf{x}; \boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*)}{\sum_{j=1}^{K} \pi_j^* \varphi_D(\mathbf{x}; \boldsymbol{c}_j^*, \boldsymbol{\Gamma}_j^*)}, \text{ for } k = 1, \dots, K.$$

We name this novel method BLLiM for *Block diagonal covariance for Gaussian Locally Linear Mapping*.

Estimation through an EM algorithm For a fixed block-diagonal structure described by $\mathbf{B} = (B_1, \ldots, B_K)$ and a fixed number of clusters K, we consider the maximum likelihood estimation for the parameters. The framework introduced above has the advantage that estimation of parameters $(\boldsymbol{c}_k, \boldsymbol{\Gamma}_k, \boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k(B_k))_{1\leq k\leq K}$ is tractable by an Expectation-Maximization (EM) algorithm, introduced in Dempster et al. (1977). Details are provided in the main paper.

Construction of the collection of models As BLLiM involves estimating both the number of clusters K and the network structure of covariates defined by $\mathbf{B} = ((d_k^{[g]})_{1\leq g\leq G_k})_{1\leq k\leq K}$, we reformulate the problem as a model selection issue. The number of clusters K varies within a candidate set $\mathcal{K}$, bounded by $K_{\max}$, so we focus on a finite set $\mathcal{K} \subset \{1, 2, \ldots, K_{\max}\}$.

The network structure $\mathbf{B}$ varies within a candidate set $\mathcal{B}$, representing partitions of the covariates indexed by $\{1, \ldots, D\}$ for each cluster of individuals. Due to the large cardinality of $\mathcal{B}$ (Bell number), we consider a moderately size subcollection $\widetilde{\mathcal{B}}$. We use hierarchical clustering with single linkage on the empirical correlation matrix of predictors, computed separately for each cluster, resulting in at most D models per cluster, ranging from one set with D variables to D sets with singletons.

This construction of model collection occurs once during initialization, allowing approximation of the block-diagonal structure without re-estimating the model collection at each EM algorithm step, which drastically reduces the computation time.

Model selection Varying $K \in \mathcal{K}$ the number of clusters and $\mathbf{B} \in \widetilde{\mathcal{B}}$ the block structure, we select a model using a penalized likelihood criterion of the form:

$$(\hat{K}, \hat{\mathbf{B}}) = \underset{K\in\mathcal{K}, \mathbf{B}\in\widetilde{\mathcal{B}}}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(f_{(K,\mathbf{B})}(\mathbf{x}_i|\mathbf{y}_i)) + \kappa\Delta_{(K,\mathbf{B})} \right\};$$

where $f_{(K,\mathbf{B})}(.|.)$ is the likelihood of BLLiM model as follows:

$$f_{(K,\mathbf{B})}(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^{K} \frac{\pi_k \varphi_L(\mathbf{y}; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)}{\sum_{j=1}^{K} \pi_j \varphi_L(\mathbf{y}; \boldsymbol{c}_j, \boldsymbol{\Gamma}_j)} \varphi_D(\mathbf{x}; \boldsymbol{A}_k\mathbf{y} + \boldsymbol{b}_k, \boldsymbol{\Sigma}_k(B_k));$$

and $\Delta_{(K,\mathbf{B})}$ is the number of parameters of the model such as:

$$\Delta_{(K,\mathbf{B})} = K\left(L + \frac{L(L+1)}{2} + D(L+1) + 1\right) + \sum_{k=1}^{K}\sum_{g=1}^{G_k} \frac{\#\{d_k^{[g]}\}(\#\{d_k^{[g]}\} - 1)}{2} - 1.$$

The classical AIC (Akaike, 1974) uses $\kappa = 2$, BIC (Schwarz, 1978) sets $\kappa = \log(n)$, and the slope heuristic (Birgé and Massart, 2001; Arlot, 2019) proposes to infer it in a data-driven manner. The slope heuristic is particularly well-suited for high-dimensional contexts, because it has theoretical non-asymptotic guarantees and because κ is adapting with respect to the data.

However, this optimization problem is costly, as we have to test every combination of $K \in \mathcal{K}, \mathbf{B} \in \widetilde{\mathcal{B}}$. We propose in this section to decompose it in a nested way, with coefficients $\kappa_{\mathbf{B}}$ and κ_K potentially different. First, K is fixed and the block structure is selected using slope heuristics:

$$\hat{\mathbf{B}}_K = \underset{\mathbf{B}\in\widetilde{\mathcal{B}}}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(f_{(K,\mathbf{B})}(\mathbf{x}_i|\mathbf{y}_i)) + \kappa_{\mathbf{B}}\Delta_{(K,\mathbf{B})} \right\} \text{ for each } K \in \mathcal{K}.$$

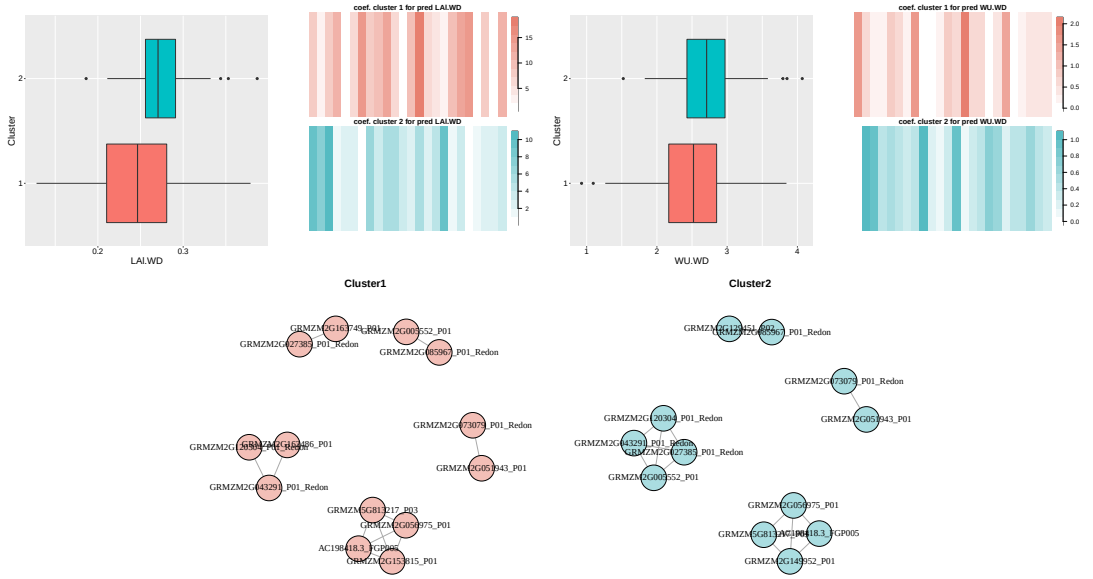
At last, the number of clusters K is estimated in the same manner:

$$\hat{K} = \underset{K\in\mathcal{K}}{\operatorname{argmin}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log(f_{(K,\hat{\mathbf{B}}_K)}(\mathbf{x}_i|\mathbf{y}_i)) + \kappa_K\Delta_{(K,\hat{\mathbf{B}}_K)} \right\}.$$

Table 2.3: Errors computed by 10-fold cross validation for the prediction of ecophysiological traits late leaf area (LAl) and water use (WU) (Prado et al., 2018) from the proteomic data from Blein-Nicolas et al. (2020). Mean and standard deviation are computed over the cross validation process.

	LAl.WD	WU.WD
mean	0.00193 (0.00322)	0.264 (0.427)
MARS (Friedman, 1991)	0.00163 (0.00364)	0.363 (0.619)
GLLiM (Deleforge et al., 2015)	0.00090 (0.00148)	0.146 (0.195)
BLLiM	0.00077 (0.00130)	0.128 (0.191)
PLS	0.00083 (0.00160)	0.121 (0.165)
RF	0.00090 (0.00157)	0.151 (0.229)
SVM	0.00090 (0.00171)	0.122 (0.166)

Figure 2.1: Response for each cluster and absolute value of the regression coefficients (top). Each color represents a cluster (red for cluster 1, blue for cluster 2). Modules of proteins (nodes) detected for each cluster of individuals (bottom). The name of the protein is written on each node. The edges correspond to interactions between proteins.



Experimental validation: application on drought-related traits in maize

We illustrate our model's performance and interpretability on the genetic and molecular bases of maize drought-responsive traits (Prado et al., 2018) and leaf proteins (Blein-Nicolas et al., 2020). This study includes 254 genotypes of dent maize, grown under two watering conditions, and phenotyped for seven ecophysiological traits (Prado et al., 2018). Leaf samples from the same plants were analyzed by proteomics, quantifying 2055 proteins (973 continuous and 1082 counting data) (Blein-Nicolas et al., 2020). For our analysis, we focused on two ecophysiological traits (LAl and WU) and the continuous protein data under water deficit conditions. After removing missing data, our dataset consisted of 233 maize genotypes ($n = 233$) with measurements for both traits ($L = 2$) and proteins ($D = 973$).

We applied the BLLiM procedure, testing clusters $\mathcal{K} = \{1, 2, 3\}$. Increasing clusters led to empty ones, so we used BIC for selection instead of the slope heuristic due to insufficient data points for calibration. For initialization, we reduced analysis to proteins selected by Lasso within each cluster, resulting in $D = 24$ variables, addressing the high-dimensionality of biological data. Our prediction strategy was compared with MARS, RF, SVM, and PLS using

10-fold cross-validation, with the mean as a baseline error measure. Predictive performance was evaluated by RMSE for each response.

Results in Table 2.3 show that only MARS performed worse than or similar to the mean. Model-free methods like RF and SVM performed well, adapting to the dataset's structure. Our procedure achieved competitive prediction performance, similar to PLS, RF, and SVM. The main drawback of model-free methods is their lack of interpretability, but the inferred model by BLLiM is interpretable.

To approximate the nonlinear relationship between the phenotypic variable and the proteomic data, BLLiM divides the individuals into two clusters of sizes 103 and 130. The response values and regression coefficients are shown in Figure 2.1. There is no clear difference in the mean levels of the responses between the clusters. The key difference lies in the link between proteins and ecophysiological traits, as indicated by the regression coefficients. For Cluster 1, proteins with high coefficients in predicting late leaf area (LAl.WD) are mainly associated with heat shock and stress responses (e.g., GRMZM5G813217, GRMZM2G153815, GRMZM2G112165, GRMZM2G043291). These proteins have small coefficients for Cluster 2. The partial correlations between proteins are also displayed in Figure 2.1. The two graphs are similar, but in Cluster 1, heat shock response proteins GRMZM5G813217 and GRMZM2G153815 are connected, whereas they are not connected in Cluster 2. These findings suggest that under water deficit, the genotypes can be distinguished by how stress response proteins relate to drought-related traits, indicating different drought response strategies.

2.2.2 NASH's prediction using mixture of logistic regression models

This project studies blood serum spectra from 395 morbidly obese patients, including 66 with NASH, aiming to develop a statistical learning model for scoring each spectrum.

Our approach emphasizes the importance of accounting for individual variability, recognizing that metabolisms vary greatly due to lifestyle, diet, and medical history. Instead of fitting patients to a rigid model, we decompose the cohort into reference profiles, or disease trajectories (Ross and Dy, 2013), which summarize metabolic behaviors and provide valuable, interpretable insights for diagnosis and treatment. We propose an intermediate approach where the joint distribution of predictors and responses is modeled as a mixture, leveraging both clustering in conditional distributions and predictor information. This approach allows for direct calculation of posterior probabilities for new observations, independent of unobserved responses. For the specific case of NASH disease data, a mixture of logistic regression models is considered, assuming Gaussian covariates and utilizing the Expectation-Maximization (EM) algorithm for inference with latent variables. Additionally, this project emphasizes the importance of sparse regression coefficients and variable selection, particularly using an ℓ_1-penalized likelihood approach. This method not only aids in accurate parameter estimation but also benefits from theoretical guarantees within the mixture of regression framework (Khalili and Chen, 2007; Städler et al., 2010). Furthermore, the graphical lasso estimator (Friedman et al., 2008) is employed to enhance precision matrix estimation, highlighting covariate dependencies while reducing dimensionality.

Overall, the proposed method aims to estimate patient profiles and interpret molecular variables affecting NASH disease, leveraging advanced statistical techniques to derive meaningful insights from complex biological data.

Penalized mixture of logistic regressions model with random design

We study $(\mathbf{X}, Y) \in \mathbb{R}^p \times \{0,1\}$ where Y is binary and $\mathbf{X}$ consists of p quantitative covariates in $\mathbb{R}^p$. The conditional relationship $f(y|\mathbf{X} = \mathbf{x})$ between Y and $\mathbf{X}$ depends on a latent class variable $\mathbf{Z} = (Z_1, \ldots, Z_K)$ following a multinomial distribution $\mathcal{M}(1, \pi_1, \cdots, \pi_K)$. Conditioned on $\mathbf{Z}$, $\mathbf{X}$ follows a Gaussian distribution, and Y is modeled using logistic regression. Parameters $\boldsymbol{\mu}_k$ and Σ_k, and $\boldsymbol{\beta}_k$ describe the mean and the covariance for each cluster k. Moreover, given $\{Z_k = 1\}$, the covariates $\mathbf{X}$ are related to the response variable Y through the logistic link function such as

$$\text{logit}\big(p^{(k)}(\mathbf{X})\big) = \mathbf{X}^T\boldsymbol{\beta}_k,$$

where logit : $x \mapsto \log(x/(1-x))$ and $\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,p})$ are the regression coefficients of the generalized linear model in cluster k.

The marginal joint distribution of $(\mathbf{X}, Y)$ (unconditional on $\mathbf{Z}$) is defined as a mixture of logistic regression with a random design. In particular, the density is given by:

$$f_{Y,\mathbf{X}}(y, \mathbf{x}) = \sum_{k=1}^{K} \pi_k f_{Y|\{\mathbf{X},\mathbf{Z}\}}(y; \boldsymbol{\beta}_k) f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k). \tag{2.6}$$

The parameters for this mixture density, denoted as $\Phi_K = (\pi_1, ..., \pi_K, \boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_K)$, include π_k and $\boldsymbol{\phi}_k = (\boldsymbol{\mu}_k, \Sigma_k, \boldsymbol{\beta}_k)$ for each cluster k.

This model, also introduced by Xu et al. (1995), is distinct from traditional finite mixture regression models (see *e.g* Grün and Leisch, 2007; Khalili and Chen, 2007) due to its random design framework. This framework allows for inference of cluster membership probabilities (Hoshikawa, 2013) based on observed covariates and facilitates out-of-sample prediction, addressing issues like optimism bias in fixed-design regressions (Rosset and Tibshirani, 2019). For a new observation $\mathbf{x}_0$, the prediction rule is:

$$\mathbb{E}(Y_0|\mathbf{X}_0 = \mathbf{x}_0) = \sum_{k=1}^{K} \mathbb{P}(Y_0 = 1|Z_{0k} = 1, \mathbf{X}_0 = \mathbf{x}_0)\mathbb{P}(Z_{0k} = 1|\mathbf{X}_0 = \mathbf{x}_0),$$

where $\mathbf{Z}_0 = (Z_{01}, ..., Z_{0K})$ is the latent variable associated with $\mathbf{X}_0$. Replacing both quantities with their expressions in the joint mixture of logistic regressions framework, a predicted value is given by

$$\widehat{Y_0} = \mathbb{E}(Y_0|\mathbf{X}_0 = \mathbf{x}_0) = \sum_{k=1}^{K} \tau'_{0,k} \frac{\exp(\mathbf{x}_0^t \boldsymbol{\beta}_k)}{1 + \exp(\mathbf{x}_0^t \boldsymbol{\beta}_k)},$$

$$\text{where } \tau'_{0,k} = \mathbb{P}(Z_{0k} = 1|\mathbf{X}_0 = \mathbf{x}_0) = \frac{\pi_k f_{\mathcal{N}}(\mathbf{x}_0; \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{\ell=1}^{K} \pi_\ell f_{\mathcal{N}}(\mathbf{x}_0; \boldsymbol{\mu}_\ell, \Sigma_\ell)},$$

with $f_{\mathcal{N}}(\,\cdot\,; \boldsymbol{\mu}, \Sigma)$ being the multivariate Gaussian density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

Model (2.6) is akin to mixture of experts (MoE) models, as discussed in literature such as Jacobs et al. (1991) and Yuksel et al. (2012), which emphasize prediction within a mixture framework. In our model, the latent cluster variable $\mathbf{Z}$ follows a multinomial distribution, influencing posterior probabilities as weights dependent on covariates. This can be viewed as a specific instance of MoE using Gaussian forms in the gating mechanism (Yuksel et al., 2012).

Given a sample $(\mathbf{x}_i, y_i)_{i=1,...,n}$ of n independent realizations of the random variables $(\mathbf{X}, Y)$, the unknown parameters $\Phi_K = (\pi_1, ..., \pi_K, \boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_K)$ are estimated by maximizing the likelihood. Our work focuses on moderate-dimensional covariates $\mathbf{X}$, where selecting relevant variables for predicting Y and estimating unstructured covariance matrices is challenging. We use a penalized likelihood method with dual penalties: a Lasso penalty for clusterwise feature selection in logistic regressions, and a Graphical Lasso penalty (Friedman et al., 2008) for controlling covariance matrix estimation in clustering. Lasso is chosen for its low generalization error in generalized linear models (Tibshirani, 1996), although exploring alternative penalties is a potential future direction. This leads to, for $\lambda_k \geq 0, \rho_k \geq 0$, for all $k = 1, \ldots, K$,

$$\ln \mathcal{L}(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n; \Phi_K) - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1 - \sum_{k=1}^{K} \rho_k \|\Theta_k\|_1, \tag{2.7}$$

where $\|\boldsymbol{\beta}_k\|_1 = \sum_{j=1}^{p} |\beta_{k,j}|$, $\Theta_k = \Sigma_k^{-1}$ is the precision matrix in the kth cluster and $\|\Theta_k\|_1$ denotes the sum of the absolute values of the elements of Θ_k. The regularization constants λ_k and ρ_k drive the amount of shrinkage on the parameters $\boldsymbol{\beta}_k$ and Θ_k for every cluster k, $k = 1, \ldots, K$. No structure is assumed within clusters, the regression coefficients and the precision matrices may have different supports in each cluster. Moreover, variables that are correlated conditionally to the others (encoded through Θ) in the Gaussian model are not necessarily useful to predict the response Y, and those useful to predict Y are not necessarily correlated, so no common structure is assumed.

Penalized maximum likelihood estimation consists in maximizing the convex function (2.7) with respect to parameters Φ_K. In a latent variable framework, this objective is usually achieved through an Expectation-Maximisation (EM) algorithm (Dempster et al., 1977). Details are provided in the main paper.

The tuning parameters $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_K)$ and $\boldsymbol{\rho} = (\rho_1, ..., \rho_K)$ control the regularization in penalized likelihood approaches, balancing between model complexity and data fit. Methods like Generalized Cross-Validation (GCV, Fan and Li (2001); Khalili and Chen (2007)) and Bayesian Information Criterion (BIC, Schwarz (1978)) are used to select these parameters. While GCV can be computationally intensive and prone to selecting irrelevant variables (Wang et al., 2007), BIC offers a trade-off by penalizing the number of parameters in the model (for example Wang et al., 2007; Khalili and Lin, 2013; Jiang et al., 2018; Lloyd-Jones et al., 2018). For a given number of clusters K, the BIC is defined as

$$BIC^{(\boldsymbol{\lambda},\boldsymbol{\rho})} = -2 \ln \mathcal{L}\left(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n; \hat{\Phi}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}\right) + \nu^{(\boldsymbol{\lambda},\boldsymbol{\rho})} \ln(n),$$

with $\hat{\Phi}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}$ the arguments of the maximum of the penalized log-likelihood function with tuning parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$. The quantity $\nu^{(\boldsymbol{\lambda},\boldsymbol{\rho})}$ counts the number of free parameters, corresponding to the number of non-zero coefficients of the model.

The number of clusters K is a sensible parameter because it describes the heterogeneity of the population. In an unsupervised setting, K is unknown and thus has to be estimated as well. Besides variable selection, the BIC is also commonly used to determine the number of clusters K in a mixture models framework (Keribin, 2000). For a given number of clusters K, the BIC is defined as

$$BIC_K = -2 \ln \mathcal{L}(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n; \hat{\hat{\Phi}}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}) + \nu_K \ln(n),$$

with $\hat{\hat{\Phi}}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}$ the maximum likelihood estimator restricted to the relevant variables and ν_K the number of free parameters of the model estimated with K clusters. However, for model-based clustering, Integrated Classification Likelihood (ICL, Biernacki et al. (2000)) is preferred over BIC due to its incorporation of cluster separation via an entropy term. Finally, for predictive purpose, the Akaike Information Criterion (AIC) is known to be more suitable (Shmueli, 2010). With the previous notations, the AIC is defined as

$$AIC_K = -2 \ln \mathcal{L}(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n; \hat{\hat{\Phi}}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}) + 2\nu_K.$$

Experimental validation: the NASH data set

The dataset consists of 395 patients, including 66 with NASH (approximately 17%), from the Nice hospital in France. It includes clinical variables and spectrometric measures from sera samples. Spectrometric curves, reflecting the metabolic profile influenced by liver condition, serve as molecular fingerprints. Experts selected specific curve portions relevant to metabolic variations linked to liver conditions, known as spectral variables. The prediction model primarily uses these spectral variables, while biological and clinical variables aid interpretation.

A model with 2 clusters was selected according to the AIC value. The proportions of each cluster are 0.66 and 0.34. The proportion of diseased patients changes according to the cluster: 19 % in cluster 1 and 12 % in cluster 2.

Table 2.4 shows the prediction performance of various models, with PMLR without clusters (denoted PLR) as the baseline. The PLMR-2 model, with two clusters, achieves the highest AUROC (0.75) and classification rate (0.76), supported by the lowest AIC and BIC values, indicating consistency between model selection metrics and cross-validation results. Its high negative predictive value suggests effective screening capabilities. Comparisons with PMLR-HA-2 (hard assignment) and PMLR-2-diag (diagonal covariance matrices) show that while hard assignment achieves similar AUROC and sensitivity, fuzzy assignment generally results in higher specificity and classification rates. Sparse covariance matrices are crucial for optimal performance, highlighting the complexity of modeling this dataset accurately.

Figure 2.2(left) illustrates the distribution of predicted scores relative to true class labels in the validation set. The threshold for classifying NASH patients, chosen to balance sensitivity

	PMLR-2	PMLR-HA-2	PMLR-2-diag	PMLR-3	PLR	LR
AUROC	**0.75**	0.75	0.69	0.68	0.64	0.67
Se	0.77	0.77	0.77	**0.85**	0.62	0.69
Sp	**0.76**	0.58	0.64	0.5	0.62	0.7
NPV	**0.94**	0.93	0.93	**0.94**	0.89	0.92
PPV	**0.38**	0.26	0.29	0.25	0.24	0.31
CR	**0.76**	0.61	0.66	0.56	0.62	0.7

Table 2.4: Comparison of prediction performance across various methods applied to the NASH dataset, evaluating metrics such as Area Under the Receiver Operating Characteristic (AUROC), sensitivity (Se), specificity (Sp), negative predictive value (NPV), positive predictive value (PPV), and classification rate (CR).

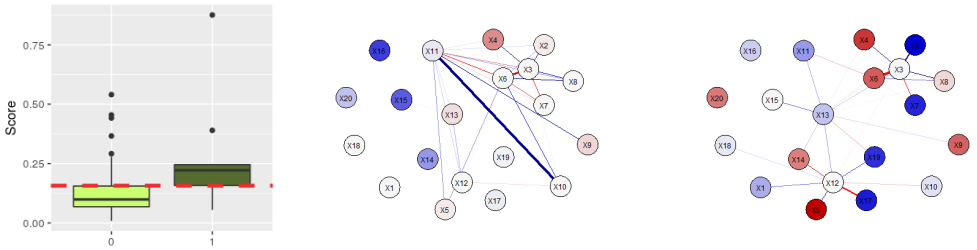


Figure 2.2: Left: Predicted scores from Penalized Mixture of Logistic Regression with 2 clusters (PMLR) plotted against true class labels in the validation set. The red dashed line represents the automatically determined threshold for classification. Middle and Right: Graphical models depicting cluster-specific relationships. Arrow colors indicate partial correlation (blue for positive, red for negative) and edge widths reflect correlation strength. Node colors denote regression coefficient values (blue for positive, red for negative) with intensity indicating magnitude, while uncolored nodes signify coefficients of zero (irrelevant variables).

and specificity, is marked by a red dashed line, effectively separating NASH patients from controls.

The graphical models derived from sparsely estimated precision matrices for each cluster are shown in Figure 2.2(middle and right). Variables that are conditionally correlated in the Gaussian model (encoded through Θ) are not necessarily predictive of the response variable Y. For example, in Cluster 1, variables $\beta_{1,15}$ and $\beta_{1,16}$ have non-zero coefficients while $[\Theta_1]15,16 = 0$, and $\beta_{1,10}$ and $\beta_{1,11}$ are zero despite $[\Theta_1]_{10,11} \neq 0$. In Cluster 1, variables X2 to X11 form a dense group, while in Cluster 2, two distinct groups are observed: X2, X3, X4, X6, X7, and X8, and X1, X5, X12, X14, X17, and X19. These varying link patterns highlight different metabolic mechanisms among patients. Node colors represent the coefficient values of each variable. Cluster 1 has a sparse model with many coefficients near zero, while Cluster 2 shows more extreme values, underscoring distinct metabolic profiles in each patient cluster.

The clusters identified by the selected model (PMLR-2) are characterized using clinical variables examined *post hoc* for interpretation. Analyzing variable distributions across clusters (details in the main paper) reveals distinct patterns in variables. Cluster 1, associated with higher diabetes indicators and liver markers, suggests more severe liver complications compared to Cluster 2. Importantly, there is no significant disparity in morphological variables (weight, height, BMI), demonstrating the model's ability to distinguish liver injury severity even among physically similar patients.

Chapter 3

Causal inference for time series

This chapter is the result of a close collaboration with Eric Gaussier (LIG), and collaborations with Julyan Arbel (LJK), Gregor Goessler (LIG), Wilfried Thuillier (LECA), and the supervision of Charles K. Assaad (PhD student), Daria Bystrova (postdoc student), Anouar Meynaoui (postdoc student), Lei Zan (PhD student). The Python code associated to the methods has been developed by Charles K. Assaad[a] and Lei Zan[b]. Thanks to them!

- Identifiability of total effects from abstractions of time series causal graphs, C.K. Assaad, E. Devijver, E. Gaussier, G. Goessler, A. Meynaoui, UAI 2024, link arxiv.
- Causal Discovery from Time Series with Hybrids of Constraint-Based and Noise-Based Algorithms, D. Bystrova, C.K. Assaad, J. Arbel, E. Devijver, E. Gaussier, W. Thuillier, TMLR, 2024, link.
- A Conditional Mutual Information Estimator for Mixed Data and an Associated Conditional Independence Test, L. Zan, A. Meynaoui, C.K. Assaad, E. Devijver and E. Gaussier, Entropy 2022, 24(9), link.
- Entropy-Based Discovery of Summary Causal Graphs in Time Series, C.K. Assaad, E. Devijver and E. Gaussier, Entropy 2022, 24(8), link.
- Survey and Evaluation of Causal Discovery Methods for Time Series , C.K. Assaad, E. Devijver and E. Gaussier, Journal of Artificial Intelligence Research, 73, 2022, link.
- Discovery of Extended Summary Graphs in Time Series, C.K. Assaad, E. Devijver and E. Gaussier, UAI 2022, link.
- A Mixed Noise and constraint-based Approach to Causal Inference in Time Series, K. Assaad, E. Devijver and E. Gaussier, ECML PKDD 2021, link.

[a] and is available at https://github.com/ckassaad
[b] and is available at https://github.com/leizan/CMIh2022

Causality plays a central role in science and has captivated the attention of philosophers, biologists, mathematicians and physicists, to name but a few. Embedded within the fabric of language and our cognitive mechanisms, causality prompts us to inquire about the nature of the world around us. Questions such as "Why is it dark?" or "What is the effect of exercise on heart rate?" reflect our innate inclination to understand causal relationships. As Spirtes, Glymour, and Scheines aptly argue, both the baby and the seasoned scientist endeavor to transform observations into causal knowledge Spirtes et al. (2001).

Causality is indeed crucial for explanatory purposes, as it allows us to understand how effects are driven by their causes, regardless of any correlations they may have with other variables. Over the recent decades, experts from various domains, including philosophy, mathematics, and computer science, have developed diverse models and methods to uncover causal relationships from data. These advancements have empowered researchers to analyze causal links effectively, leading to valuable insights and applications in diverse fields.

While the initial focus was on inferring causal relations from non-temporal data, there has been a notable shift towards analyzing time series data in recent years. This shift has spurred the development of tailored techniques for temporal datasets, enabling researchers to address complex real-world challenges. From healthcare to industrial applications, the ability to predict outcomes, such as the effects of interventions, has become increasingly essential, underscoring the growing significance of causal inference in contemporary research.

This chapter diverges somewhat from the format of previous chapters, as its content has been restructured to enhance readability, minimize redundancy, and emphasize our contributions. The chapter stems from the CIFRE PhD theses of Charles Assaad and Lei Zan, which I co-supervised with Eric Gaussier and with the company Easyvista, along with the postdoctoral work of Anouar Meynaoui, conducted in collaboration with Charles Assaad, Eric Gaussier, and Gregor Goessler. Additionally, an auxilary project from Charles's PhD, conducted in collaboration with Daria Bystrova, also a PhD student at that time under the supervision of Julyan Arbel and Wilfried Thuillier, is presented in this chapter.
The organization of this chapter is the following:

- In Section 3.1, we provide background information to understand the problem of causal inference for time series. We begin by outlining classical assumptions, delve into causal discovery using constraint-based methods, discuss the necessary independence measures, and explore the types of causal graphs relevant to time series analysis.

- In Section 3.2, we propose some contributions to estimate and test independence based on mutual information. Particularly, Section 3.2.1 focuses on mixed data, a practical but under-theorized area addressed in Lei Zan's PhD thesis detailed in Zan et al. (2022). Section 3.2.2 covers several independence measures tailored for time series, relevant to Charles Assaad's PhD thesis elaborated in Assaad et al. (2022a,b).

- In Section 3.3, we propose some contributions on causal discovery for time series. Building on previously introduced independence measures, we adapt the PC algorithm, a standard constraint-based method, to accommodate time series causal graphs. This work is also detailed in Charles Assaad's PhD thesis Assaad et al. (2022a,b, 2021). We also discuss the literature, presented in details in a survey (Assaad et al., 2022c), and present some experimental results.

- In Section 3.4, we present a contribution on the identifiability of total effects in time series causal graphs. Specifically, we address scenarios where only partial knowledge is available to identify queries at specific timestamps. This corresponds to Anouar Meynaoui's postdoctoral research, and is detailed in Assaad et al. (2024).

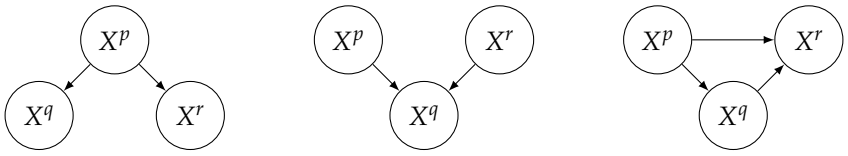


Figure 3.1: A confounder (left), a collider (middle) and a mediator (right). .

3.1 Background

3.1.1 Classical assumptions

Causal inference aims to construct a causal graph from observational data, where understanding the relationship between a probability distribution and its graphical representation is pivotal. However, given the maxim "correlation is not causation," additional assumptions are necessary to discern causal relations.

Firstly, let's examine three fundamental causal structures depicted in Figure 3.1. The left structure represents a confounder—a variable serving as a common cause for two others. In the middle, we have a collider—a variable influenced by two unrelated factors. On the right is a mediator. Without observing the common cause X^p in the confounder structure, one might infer a spurious correlation and a causal link between X^q and X^r, as these variables are independent only when conditioned on X^p. To mitigate such spurious correlations, one approach is to assume the measurement of all common causes.

Definition 3.1.1 (Causal Sufficiency, Spirtes et al. (2001)[1])**.** *A set of variables is said to be causally sufficient if all common causes of all variables are observed.*

Under the assumption of causal sufficiency, most of the causal discovery algorithms assume that the causal structure can be represented by a Directed Acyclic Graph (DAG) where directed edges signify relationships from causes to effects. The absence of an edge between two variables implies their (conditional) independence. Whenever a probability distribution can be factorized according to a given DAG, we say that the DAG and the probability distribution are compatible. The relationship between the (conditional) independence or dependence of variables and the topology of the graph, given compatible graphs and probability distributions, is grounded in the concept of d-separation, initially introduced within the framework of Bayesian networks.

Definition 3.1.2 (d-separation, Pearl (1988))**.** *If $\mathcal{G}$ is a DAG in which X^p and X^q are two vertices and X^R is a set of vertices, then X^p and X^q are d-connected by X^R in $\mathcal{G}$ if and only if there exists an undirected path U between X^p and X^q such that for every collider X^c on U, either X^c or a descendant of X^c is in X^R, and no non-collider on U is in X^R. Otherwise, X^p and X^q are d-separated given X^R.*

The following theorem states a necessary and sufficient condition for a DAG and a probability distribution to be compatible.

Theorem 3.1.3 (Markov Condition, Pearl (2000))**.** *A necessary and sufficient condition for a probability distribution to be compatible with a DAG $\mathcal{G}$ is that every variable be independent of all its nondescendants (in $\mathcal{G}$), conditional on its parents.*

When interpreting the DAG causally, the parents of a variable correspond to its direct causes, known as the Causal Markov Condition Spirtes et al. (2001). It's important to note that several DAGs can represent the same set of conditional independencies and be compatible with the same probability distribution. To address this challenge, two additional conditions have been introduced to constrain the graphs considered for a given probability distribution. The first condition is the minimality condition, which mandates that the graph does not contain dependencies absent in the observational data.

Definition 3.1.4 (Minimality Condition, Pearl (2000))**.** *A DAG $\mathcal{G}$ compatible with a probability distribution P is said to satisfy the minimality condition if P is not compatible with any proper subgraph of $\mathcal{G}$.*

[1] All this chapter is restricted to this context, even though I did some work to relax this assumption (mainly based on extending FCI Spirtes et al. (2001); Zhang (2008); Colombo et al. (2012)).

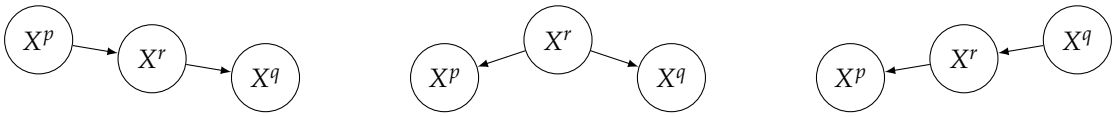


Figure 3.2: Three equivalent structures

The minimality condition is however not sufficient to restrict the set of possible causal structures, and we introduce the faithfulness assumption.

Definition 3.1.5 (Faithfulness, Spirtes et al. (2001))**.** *We say that a graph $\mathcal{G}$ and a compatible probability distribution P are faithful to one another if all and only the conditional independence relations true in P are entailed by the Markov condition applied to $\mathcal{G}$.*

Note that the minimality condition is weaker than faithfulness in the sense that faithfulness and Markov conditions together entail minimality, whereas both minimality and Markov conditions do not always entail faithfulness.

3.1.2 Causal discovery with constraint-based methods

Constraint-based approaches exploit conditional independencies to construct a skeleton between variables. This skeleton is then oriented according to a set of rules that define constraints on admissible orientations. Central to these approaches is the notion of v-structures (Figure 3.1 (right)), or colliders, as these are the only structures which can be oriented without ambiguity.

Under causal sufficiency (Assumption 3.1.1), the underlying causal graph is typically represented by a DAG, but multiple DAGs can encode the same set of conditional independencies. For example, the models in Figure 3.2, borrowed from Verma and Pearl (1991), represent the same independence relation: $X^p \perp\!\!\!\perp X^q | X^r$. This leads to the concept of Markov equivalence class which corresponds to a set of DAGs that encode the same set of conditional independencies. Verma and Pearl (1991) have shown that two DAGs are Markov equivalent if and only if they have the same skeleton and v-structures. Within an equivalence class of DAGs, Andersson et al. (1997); Chickering (2002) introduced the completed PDAG (CPDAG) as the representation consisting of directed edges for every compelled edges (those participating in v-structures or potentially forming new v-structures upon orientation changes), and undirected edges for all other. A CPDAG uniquely represents a Markov equivalence class, making the goal of constraint-based algorithms clear: construct the CPDAG from observational data representing the Markov equivalence class of the true causal graph.

One of the earliest constraint-based algorithm is the SGS algorithm Spirtes et al. (2001), suffers from impracticality due to exponential growth in the number of conditional independencies to be tested, especially challenging given the difficulty in computing such independencies Shah and Peters (2020). Addressing this, the Peter-Clark (PC) algorithm Spirtes et al. (2001) was introduced. Starting with a complete undirected graph $\mathcal{G}$, PC algorithm evaluates dependencies between all pairs of vertices, iteratively removing or retaining links based on their independence status. It then assesses conditional independencies between dependent vertices, progressively increasing the number of variables to condition on until a conditional independence is found or all relevant sets of vertices have been considered. Once the skeleton has been constructed, the algorithm applies series of rules Spirtes et al. (2001); Colombo and Maathuis (2014), starting by identifying v-structures using the so-called origin of causality (Rule 0) and repeating Rules 1-2-3 on the remaining undirected edges.

PC-Rules

0. For every triple $X^p - X^r - X^q$ such that X^p and X^q are not adjacent and $X^r \notin Sepset(p, q)$, orient the triple as $X^p \to X^r \leftarrow X^q$.

1. In a triple $X^p \to X^q - X^r$ such that X^p and X^r are not adjacent, orient $X^q \to X^r$.

2. If there exist a direct path from X^p to X^q and an edge between X^p and X^q, orient $X^p \to X^q$.

3. Orient $X^p \to X^q$ whenever there are two paths $X^p - X^r \to X^q$ and $X^p - X^s \to X^q$.

The main weakness of the original PC algorithm is its dependency on the order of operations, making it inherently unstable. However, Colombo and Maathuis (2014) proposed a modification that measures all conditional independencies for a given cardinality before altering links in the undirected graph, thereby eliminating order dependence.

From a theoretical perspective, this algorithm is both sound (all causal relations detected by the rules are correct) and complete (all possible causal relations in the Markov equivalence class are detected by the algorithm) Meek (1995); Andersson et al. (1997) within the set of Markov equivalence graphs (Theorem 5.1 by Spirtes et al. (2001)). Its consistency has been discussed by Spirtes et al. (2001); Robins et al. (2003): while uniform consistency cannot be achieved if the model is only faithful, pointwise consistency is attainable. Kalisch and Bühlmann (2007); Zhang and Spirtes (2002) provided assumptions which render the PC-algorithm uniformly consistent, with the number of nodes and neighbors increasing in a limited way with respect to the sample size.

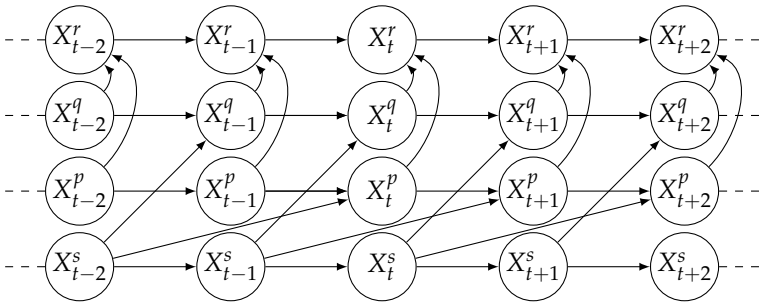
3.1.3 Independence measure: estimation and test

At the core of constraint-based algorithms are conditional (in)dependence measures, crucial for detecting relevant conditional dependencies. Numerous dependence measures have been proposed in the literature (see Josse and Holmes (2016) for a recent survey), ranging from linear models to more intricate approaches, each with distinct advantages and drawbacks, leading to a lack of universal acceptance. A key requirement is the ability to consider conditional measures, which is fundamental in causal graph analysis but poses a significant challenge in statistics. Additionally, these measures must be accompanied by statistical tests to determine if the value is significant.

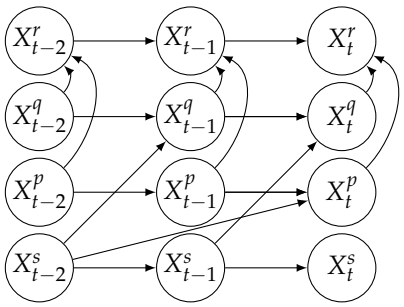
Consider first the case of independent and identically distributed (iid) variables. Conditional independence tests for categorical variables are well-established, with solutions such as the Pearson's χ^2 test and the likelihood ratio test Tsamardinos and Borboudakis (2010). However, testing testing conditional independence for continuous random variables presents greater difficulty Shah and Peters (2020). While (partial) correlation and its associated Fisher test are commonly used for linear models with Gaussian variables due to their numerical simplicity and theoretical clarity, such modeling assumptions may not hold in real-world scenarios. Nonparametric conditional independence tests, which do not assume any specific functional form between variables or data distributions, have gained popularity for their robustness. Shah and Peters (2020) show that no conditional independence test can control type-I error for all conditional independence cases, but their validity for a wide range of conditional independence cases make them the popular choice. Methods based on kernels mean embedding, such as the Hilbert-Schmidt independence criterion (HSIC, Gretton et al. (2007)) and its extension for conditional independence Fukumizu et al. (2007), as well as its refinement the Kernel Conditional Independence Test (KCIT, Zhang et al. (2011)) and its approximation by random Fourier features Strobl et al. (2019), have demonstrated validity for a wide range of scenarios. However, the computational demands of kernel-based methods remain a drawback, despite recent attempts to address this issue Jitkrittum et al. (2017); Zhang et al. (2018). Another category of measures is based on conditional distributions, such as mutual information, which has seen nonparametric testing methods developed to assess conditional independence efficiently Berrett and Samworth (2019); Berrett et al. (2020), leveraging efficient entropy estimators derived from nearest neighbor distances Berrett et al. (2019).

For time series data, adaptation of these measures and tests have been proposed, to account for lagged dependencies by considering shifted time series. Granger causality, introduced by Granger (1969), has significantly advanced our understanding of directional influence between time series. Transfer entropy (TE), pioneered by Schreiber (2000), offers an alternative to lagged mutual information, incorporating shared information from common history and input signals, and has found widespread application in various fields beyond physics. Amblard and Michel (2013) proposed a survey between Granger causality and directed information theory, and derive some links between derived measures.

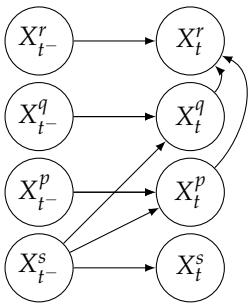
We focus in this chapter on the mutual information for analyzing time series and mixed data. To estimate entropy, two primary approaches have emerged. The first relies on kernel-



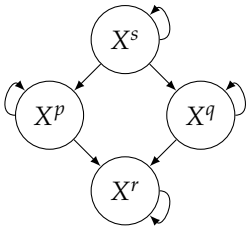
(a) Full Time Causal Graph



(b) Window Causal Graph



(c) Extended Summary Causal Graph



(d) Summary Causal Graph

Figure 3.3: Different causal graphs that one can infer from three time series.

density estimates (Beirlant et al., 1997), tailored for quantitative data, while the second utilizes on k-nearest neighbours Kozachenko and Leonenko (1987); Singh et al. (2003), suitable for both qualitative and quantitative data. The latter is preferred for its natural adaptation to data density and minimal requirement for kernel bandwidth tuning. In this method, the distance to the k^{th} nearest neighbour is computed for each data point, with the probability density around each point substituted into the entropy expression. When k is fixed and the number of points is finite, each entropy term becomes noisy, introducing bias to the estimator. However, this bias is distribution independent and can be corrected for (Singh and Póczos, 2016). Building on this approach, Kraskov et al. (2004) proposed a mutual information estimator that extends beyond the sum of entropy estimators. Subsequently, Frenzel and Pompe (2007) extended this work to conditional mutual information for quantitative data. Determining whether the estimated (conditional) mutual information value is sufficiently small to infer conditional (in)dependence typically involves statistical independence tests. Permutation tests (Berry et al., 2018) are commonly employed to avoid assumptions about data distribution. However, standard permutation tests may inadvertently disrupt the dependence structure between variables. To address this, Runge (2018b) proposed a local permutation test, preserving the dependence structure between variables by restricting permutations within similar values of a conditioning variable.

3.1.4 Causal graphs for time series

Consider a d-variate time series X where each X_t at a fixed time t is a vector $(X_t^1, \cdots, X_t^d)$, with X_t^p representing the p-th time series measurement at time t. There exist four ways (to our knowledge) for representing time series through a causal graph. The first approach, termed a full time causal graph (or infinite dynamic causal graph according to Malinsky and Spirtes (2018)), depicts a complete graph of the dynamic system (see Figure 3.3(a)).

However, inferring full time causal graphs is often impractical as there usually is a single observation for each time series at each time point. Instead, practitioners often rely on the Consistency Throughout Time assumption (also referred to as Causal Stationarity by Runge (2018a)), which states that all the causal relationships remain constant in direction throughout

time. Under this assumption, the full time causal graph can be contracted, without loss of information, to yield a finite graph which we call window causal graph, depicted in Figure 3.3(b). This representation captures causal relationships within a time window, with the window size equaling the maximum lag τ relating time series in the full time causal graph.

The window causal graph can be condensed into a summary causal graph (also referred to as a unit graph by Chu and Glymour (2008)), as depicted in Figure 3.3(d), albeit at the cost of losing information regarding the specific time points at which causation occurs. In practice, understanding causal relationships between time series as a whole is often adequate, without requiring precise knowledge of the timing of these relationships. Note that since a summary causal graph is a condensed representation of the full time causal graph, it may contain cycles.

However, in practice, even though we are not able to distinguish between different lags, it is important to distinguish between instantaneous and lagged causal relations. So we have introduced the extended summary causal graph in Assaad et al. (2022a), as depicted in Figure 3.3(c).

When considering temporal variables, the concept of temporal priority, dating back to Hume (1738), is valuable. It asserts that a cause precedes its effects, imparting an asymmetric temporal nature to the causality process. This concept aids in orienting causal relations, especially when the chronological order of events is known. However, challenges arise when the difference in time between events associated with different time series is not observed due to low sampling frequencies. This can lead to the misperception of instantaneous causal relations between events occurring at different time points in observational time series, a challenge we address in Section 3.3.

Note that for a fixed summary causal graph, there exists several extended summary causal graph that are associated, and even more different full-time causal graph that can be contracted to lead to the same summary causal graph. For a fixed abstract graph $\mathcal{G}$ (summary causal graph or extended summary causal graph), we denote $\mathcal{C}(\mathcal{G})$ the class of compatible full time causal graph.

3.2 Independence measure: some contributions

When considering mixed data, very few estimators and tests exist in the literature. We introduce in Section 3.2.1 an estimator and a related test taking into account both the qualitative and quantitative dimension of a mixed vector (this work corresponds to Zan et al. (2022)). Then we introduce several (conditional) dependence measures for time series, well-suited for summary causal graph (CTMI, introduced in Assaad et al. (2022b) and TCE, introduced in Assaad et al. (2021)) and for extended summary causal graph (GCE, introduced in Assaad et al. (2022a)).

3.2.1 Mutual information for mixed data

A standard approach to estimating (conditional) mutual information from mixed data involves discretizing the data and approximating the distribution of the random variables with a histogram model defined on a set of intervals called bins (Scott, 2015b). Each bin represents a single point for qualitative variables and consecutive non-overlapping intervals for quantitative variables. Although smaller bins improve the approximation, the finite sample size necessitates careful selection of the number of bins. To efficiently generate adaptive histogram models from quantitative variables, Cabeli et al. (2020) and Marx et al. (2021) transform the problem into a model selection problem, using a criterion based on the minimum description length (MDL) principle.

More recently, Ross (2014) and Gao et al. (2017) introduced two approaches to estimate mutual information for mixed data, however without any conditioning set. Following these studies, Rahimzamani et al. (2018) proposed a measure of incompatibility between the joint probability and its factorization called graph divergence measure and extended the estimator proposed in Gao et al. (2017) to conditional mutual information, leading to a method called RAVK. As ties can occur with a non zero probability in mixed data, the number of neighbours has to be carefully chosen. Mesner and Shalizi (2020) extended FP (Frenzel and Pompe, 2007) to the mixed data case by introducing a qualitative distance metric for non-quantitative variables, leading to a method called MS. The choice of the qualitative and quantitative distances is a crucial point in MS (Ahmad and Khan, 2019).

Hybrid conditional mutual information estimation Let us consider three mixed random vectors X, Y and Z, where any of their components can be either qualitative (stacked in X^ℓ, Y^ℓ, Z^ℓ) or quantitative (stacked in X^t, Y^t, Z^t). Then, the conditional mutual information can be decomposed into termed conditioned on qualitative components:

$$\begin{aligned} I(X;Y|Z) =& H(X^t, Z^t|X^\ell, Z^\ell) + H(Y^t, Z^t|Y^\ell, Z^\ell) - H(X^t, Y^t, Z^t|X^\ell, Y^\ell, Z^\ell) \\ & - H(Z^t|Z^\ell) + H(X^\ell, Z^\ell) + H(Y^\ell, Z^\ell) - H(X^\ell, Y^\ell, Z^\ell) - H(Z^\ell). \end{aligned} \tag{3.1}$$

Conditioning with qualitative variables leads to a simpler estimation.

Consider a sample of size N denoted $(x_i, y_i, z_i)_{i=1,\dots,N}$. We estimate the qualitative entropy terms of Equation (3.1) using histograms in which bins are defined by the Cartesian product of qualitative values; and for conditional entropies of Equation (3.1) of quantitative variables conditioned on qualitative variables, probabilities are estimated by their empirical versions and conditional entropies are estimated using the nearest neighbour estimator (Singh et al., 2003) on the sample points satisfying the conditioning set. Interestingly, the resulting conditional mutual information estimator is asymptotically unbiased and consistent.

In the main paper (Zan et al., 2022), we compare experimentally our estimator, denoted CMIh, with several estimators. The main competitor (Mesner and Shalizi, 2020) has a main drawback as it gives the value 0, or close to 0, to the estimator in some particular cases discussed in the main paper. Our proposed estimator does not suffer from this drawback as we do not directly compare two different types of distances, one for quantitative and one for qualitative data. The pure histogram method (Marx et al., 2021) performs well in terms of accuracy of the estimator, but its computation time is prohibitive. Our estimator, which can be seen as a trade-off between k-nearest neighbour and histogram methods, performs well, both in terms of the accuracy of the estimator and in terms of the time needed to compute this estimator.

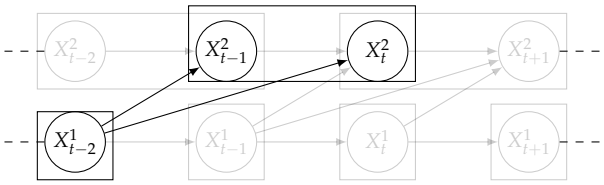


Figure 3.4: Why do we need windows and lags? An illustration with two time series where X^1 causes X^2 in two steps (circles correspond to observed points and rectangles to windows). The arrows in black are discussed in the text.

Testing conditional independence Once an estimator for mutual information has been computed, one relies on statistical tests to conclude on the dependence or independence of the involved variables, among which permutation tests are widely adopted as they do not require any modelling assumption (Berry et al., 2018). We also focus on such tests here which emulate the behaviour of the estimator under the null hypothesis (corresponding to independence) by permuting values of variables. Runge (2018b) showed that for conditional tests and purely quantitative data, local permutations that break any possible dependence between X and Y while preserving the dependence between X and Z and between Y and Z are to be preferred over global permutations. Our contribution here has been to extend this method to mixed data. Experimental results have shown that the local tests are better than the global one, and considered the mixed nature of the data is giving better performances.

3.2.2 Mutual information for time series

We present in this section some new mutual information measures whether time series are (conditionally) dependent or not. We assume that all time series are aligned in time, with the same sampling rate, but this as been relaxed in the corresponding papers. Without loss of generality, time instants are assumed to be integers.

First, the following example illustrates why this is a complex task.

Example 3.2.1. *Let us consider the following two time series defined by, for all t,*

$$X^1_t = X^1_{t-1} + \xi^1_t,$$
$$X^2_t = X^2_{t-1} + X^1_{t-2} + X^1_{t-1} + \xi^2_t,$$

with $(\xi^1_t, \xi^2_t) \sim \mathcal{N}(0,1)$. The corresponding full time causal graph is displayed in Figure 3.4. In order to capture the dependencies between the two time series, one needs to take into account a lag between them, as the true causal relations are not instantaneous, as done for example in Runge et al. (2019). A window-based representation is also necessary to fully capture the dependencies between the two time series. Indeed, as X^2_{t-1} and X^2_t are the effects of the same cause (X^1_{t-2}), it may be convenient to consider them together when assessing whether the time series are dependent or not. For example, defining (overlapping) windows of size two for X^2 and one for X^1 with a lag of 1 from X^1 to X^2, as in Figure 3.4, allows one to fully represent the causal dependencies between the two time series.

We propose several measures that will be used in Section 3.3 for causal discovery with time series. More precisely, CTMI and TCE are well-suited for inferring a summary causal graph, while GCE is distinguishing between instantaneous and lagged causal relations, being particularly well-suited for inferring an extended summary causal graph.

Causal Temporal Mutual Information Let us consider d univariate time series $X^1, \cdots, X^d$ and their observations $(X^p_t)_{1 \leq t \leq N_p, 1 \leq p \leq d}$.

Definition 3.2.1. *Let $\gamma_{\max}$ denote the maximum lag between two time series X^p and X^q, and let the maximum window size $\lambda_{\max} = \gamma_{\max} + 1$. The window-based representation, of size $0 < \lambda_{pq} \leq \lambda_{\max} < N_p$, of the time series X^p with respect to X^q, which will be denoted $X^{(p;\lambda_{pq})}$, simply amounts to considering $(N_p - \lambda_{pq} + 1)$ windows: $(X^p_t, \cdots, X^p_{t+\lambda_{pq}-1})$, $1 \leq t \leq N_p - \lambda_{pq} + 1$. The window-based representation, of size $0 < \lambda_{qp} \leq \lambda_{\max} < N_q$, of the time series X^q with respect to X^p is defined in the same way. A*

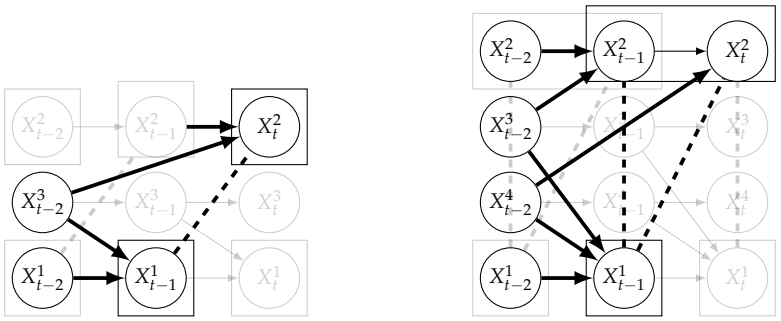


Figure 3.5: Example of conditional independence between dependent time series. Left, the conditioning set contains one time series X^3 in addition to the past of X^1 and X^2. Right, the conditioning set contains two time series X^3 and X^4 in addition to the past of X^1 and X^2. Dashed lines are correlations that are not causations, bold arrows correspond to conditioning variables.

temporal lag $\gamma_{pq} \in \mathbb{Z}$ compatible with λ_{pq} and λ_{qp} relates windows in $X^{(p;\lambda_{pq})}$ and $X^{(q;\lambda_{qp})}$ with starting time points separated by γ_{pq}. We denote by $\mathcal{C}^{(p,q)}$ the set of window sizes and compatible temporal lags.

Based on the above elements, we define the causal temporal mutual information between two time series X^p and X^q as the maximum of the standard mutual information over all possible compatible temporal lags and windows $\mathcal{C}^{(p,q)}$, conditioned by the past of the two time series.

Definition 3.2.2. *Consider two time series X^p and X^q. We define the causal temporal mutual information between X^p and X^q as:*

$$CTMI(X^p; X^q) = \max_{(\lambda_{pq},\lambda_{qp},\gamma_{pq})\in\mathcal{C}^{(p,q)}} I(X_t^{(p;\lambda_{pq})}; X_{t+\gamma_{pq}}^{(q;\lambda_{qp})} | X_{t-1}^{(p;1)}, X_{t+\gamma_{pq}-1}^{(q;1)}), \tag{3.2}$$

where I represents the mutual information. In case the maximum can be obtained with different values in $\mathcal{C}^{(p,q)}$, we first set $\bar{\gamma}_{pq}$ to its largest possible value. We then set $\bar{\lambda}_{pq}$ to its smallest possible value and, finally, $\bar{\lambda}_{qp}$ to its smallest possible value. $\bar{\gamma}_{pq}$, $\bar{\lambda}_{pq}$, and $\bar{\lambda}_{qp}$, respectively, correspond to the optimal lag and optimal windows.

Under consistency throughout time, CTMI satisfies the standard properties of mutual information, namely it is nonnegative, symmetric, and equals 0 iff time series are independent. Thus, two time series X^p and X^q such that $CTMI(X^p; X^q) > 0$ are dependent. We now extend the causal temporal mutual information by conditioning on a set of variables. Figure 3.5 illustrates two cases where the dependence between X^1 and X^2 are due to spurious correlations originating from common causes. Conditioning on these common causes should lead to the conditional independence of the two time series. This leads us to the following definition of the conditional causal temporal mutual information.

Definition 3.2.3. *Consider two time series X^p and X^q and a set $X^{\mathbf{R}} = \{X^{r_1}, \cdots, X^{r_K}\}$. We define the conditional causal temporal mutual information between X^p and X^q conditionet on $X^{\mathbf{R}}$ as:*

$$CTMI(X^p; X^q \mid X^{\mathbf{R}}) = I(X_t^{(p;\bar{\lambda}_{pq})}; X_{t+\bar{\gamma}_{pq}}^{(q;\bar{\lambda}_{qp})} | (X_{t-\bar{\Gamma}_k}^{(r_k;\bar{\lambda}_k)})_{1\leq k\leq K}, X_{t-1}^{(p;1)}, X_{t+\bar{\gamma}_{pq}-1}^{(q;1)}). \tag{3.3}$$

In case the minimum can be obtained with different values, we first set $\bar{\Gamma}_k$ to its largest possible value. We then set $\bar{\lambda}_k$ to its smallest possible value.

Temporal Causation Entropy Causation entropy, introduced in Sun et al. (2015b), is an asymmetric measure that detects the uncertainty reduction of the future states of X^q as a result of knowing the past states of X^p given that the past of $X^{\mathbf{R}}$ is already known, where $\mathbf{R}$ is a subset of $\{1, \cdots, d\}$. We extend the standard causation entropy measure to handle instantaneous relations and lags bigger than one.

Definition 3.2.4 (Temporal causation entropy)**.** *Consider two time series X^p and X^q. We first define the optimal lag γ_{pq} between X^p and X^q and (λ_{pq}, λ_{qp}) the optimal windows of X^p regarding X^q and of*

X^q regarding X^p respectively as:

$$\gamma_{pq}, \lambda_{pq}, \lambda_{qp} = \underset{\gamma \geq 0, \lambda_1, \lambda_2}{\operatorname{argmax}} \; h(X^q_{t:t+\lambda_2} \mid X^q_{t-1}, X^p_{t-\gamma-1}) - h(X^q_{t:t+\lambda_2} \mid X^p_{t-\gamma-1:t-\gamma+\lambda_1}, X^q_{t-1}),$$

where h denotes the entropy. The temporal causation entropy from X^p to X^q conditioned on a set $X^{\boldsymbol{R}} = \{X^{r_1}, \cdots, X^{r_K}\}$ is given by:

$$\begin{aligned} TCE(X^p \to X^q \mid X^{\boldsymbol{R}}) = \min_{\Gamma_{r_i} \geq 0, 1 \leq i \leq K} \; & h(X^q_{t:t+\lambda_{qp}} \mid (X^{r_i}_{t-\Gamma_{pq|r_i}})_{1 \leq i \leq K}, X^q_{t-1}, X^p_{t-\gamma_{pq}-1}) \\ & - h(X^q_{t:t+\lambda_{qp}} \mid (X^{r_i}_{t-\Gamma_{pq|r_i}})_{1 \leq i \leq K}, X^p_{t-\gamma_{pq}-1:t-\gamma_{pq}+\lambda_{pq}}, X^q_{t-1}), \end{aligned}$$

where $\Gamma_{pq|r_1}, \cdots, \Gamma_{pq|r_K}$ are the lags between $X^{\boldsymbol{R}}$ and X^q.

First, the lag between X^p and X^q is detected by maximizing the dependency between X^p and X^q. As we measure the amount of information brought by the observations of one variable on the observations of another variable, taking the maximum ensures that one does not miss any possible information contributing to relating the two time series. In a second step, we find the lags between (X^p, X^q) and $X^{\mathbf{R}}$ that minimize the conditional dependency between X^p and X^q conditioned on $X^{\mathbf{R}}$. Taking the minimum ensures that we search for the lags that break the maximal dependence. Following the temporal priority principle, which states that causes precede their effects in time, we also ensure while finding only nonnegative lags that X^p as well as the conditional variables should precede in time X^q. If $\gamma = 1$ and $\lambda_{pq} = \lambda_{qp} = 1$, then the temporal causation entropy is equivalent to causation entropy when the latter is conditioned on the past.

Compared to CTMI, TCE is asymmetric, so it already distinguishes between a cause and its effect.

Greedy causation entropy When considering extended summary causal graph, one should measure the dependence between the present and the past. The following proposition gives a characterization of this measure with the past.

Proposition 3.2.5. *Consider two time series X^p and X^q. Let γ denote the maximum gap between a cause X^p and its effect X^q. The following two propositions are equivalent:*

(a) $I(X^q_t; X^p_{t-\gamma_1}, \cdots, X^p_{t-\gamma_K}) = 0$, $\forall K \geq 1$, $\forall \gamma_1 > \cdots > \gamma_K \geq 0$,

(b) $I(X^q_t; X^p_{t-\gamma:t}) = 0$.

The same equivalence holds for the conditional mutual information, using any conditional set.

To assess whether there exist causal relations between variables in the past and potential effect in the present slices, we make use of the following greedy causation entropy[2] which is based on Prop. 3.2.5 and is asymmetric to reflect the specific role of the cause and the effect. Relations between variables in the past and present slices are naturally oriented by temporal priority.

Definition 3.2.6. *Consider two time series X^p and X^q. The greedy causation entropy, denoted by GCE, from X^p to X^q is defined by:*

$$GCE(X^p \to X^q) = I(X^q_t; X^p_{t-\gamma:t-1}).$$

Denoting by $X^{\boldsymbol{Pr}}$ a set of m time series $\{X^{Pr_1}, \cdots, X^{Pr_m}_t\}$ in the present slice and by $X^{\boldsymbol{Pa}}$ a set of ℓ time series $\{X^{Pa_1}_{t-}, \cdots, X^{Pa_\ell}_{t-}\}$ in the past slice, the conditional greedy causation entropy furthermore takes the form:

$$GCE(X^p \to X^q | X^{\boldsymbol{Pa}}, X^{\boldsymbol{Pr}}) = I(X^q_t; X^p_{t-\gamma:t-1} | X^{Pa_1}_{t-}, \cdots, X^{Pa_\ell}_{t-}, X^{Pr_1}_t, \cdots, X^{Pr_m}_t).$$

[2] We call it greedy because it considers all past instants (up to γ) without trying to filter them.

Because of Prop. 3.2.5, one can conclude that past instants of X^p do not directly cause X^q iff there exists $X^{\mathbf{Pr}} = \{X_t^{Pr_1}, \cdots, X_t^{Pr_m}\}$ and $X^{\mathbf{Pa}} = \{X_{t-}^{Pa_1}, \cdots, X_{t-}^{Pa_\ell}\}$, with $m, \ell \geq 0$, such that $\mathrm{GCE}(X^p \to X^q | X^{\mathbf{Pa}}, X^{\mathbf{Pr}}) = 0$.

When considering an extended summary causal graph, for determining (in)dependencies in the present slice, one can directly rely on the standard (conditional) mutual information.

3.3 Causal discovery for time series: some contributions

Once we have a dependence measure between time series, we can use constraint-based methods to infer a causal graph. In this section, we first detail some propositions to infer a causal graph for time series, and then provide a summary of the state-of-the-art (up to date in 2022, a part of the survey Assaad et al. (2022c) used for illustration and comparison).

3.3.1 Proposed methods

PCTMI (with ERP) When inferring a summary causal graph, we do not have to consider all the potential dependencies between two time series (which would be necessary for inferring a window causal graph). Using the maximum over all possible associations is a way to summarize all temporal dependencies, which ensures that one does not miss a dependency between the two time series. Furthermore, conditioning on the past allows one to eliminate spurious dependencies in the form of auto-correlation, as in transfer entropy (Schreiber, 2000). Interestingly, CTMI can be related to a version of the probability raising principle (Suppes, 1970), which states that a cause, here a time series, raises the probability of any of its effects, here another time series, even when the past of the two time series is taken into account, meaning that the relation between the two time series is not negligible compared to the internal dependencies of the time series.

We consider a slightly different principle based on the causal temporal mutual information, which we refer to as the entropy reduction principle. Let X^p and X^q be two time series with window sizes λ_{pq} and λ_{qp}. We say that X^p is an entropic prima facie cause of X^q with delay $\gamma_{pq} > 0$ iff $I(X_t^{(p;\lambda_{pq})}; X_{t+\gamma_{pq}}^{(q;\lambda_{qp})} | P_{t,t+\gamma_{pq}}) > 0$, which is equivalent to considering that the entropy of X^q when conditioned on the past reduces when one further conditions on X^p.

In addition to the PC orientation rules, we introduce two new rules, which are based on the notion of possible spurious correlations and the mutual information we have introduced. The notion of possible spurious correlations captures the fact that two variables may be correlated through relations that do not only correspond to direct causal relations between them (there exists a path between them that neither contains the edge $X^p - X^q$ nor any collider).

Interestingly, when two connected variables do not have possible spurious correlations, one can conclude about their orientation using CTMI.

Proposition 3.3.1. *Let us assume that we are given perfect conditional independence information about all pairs of variables (X^p, X^q) in V given subsets $S \subseteq V \backslash \{X^p, X^q\}$. Then, every non-oriented edge in the CPDAG obtained by the above procedure corresponds to a prima facie cause and by, causal sufficiency, to a true causal relation between the related time series. Furthermore, the orientation of an unoriented edge between two nodes X^p and X^q that do not have possible spurious correlations is given by the "direction" of the optimal lag in $CTMI(X^p, X^q)$, assuming that the maximal window size is larger than the longest lag γ_{max} between causes and effects.*

The following orientation rule is a direct application of the above proposition.

ER-rule 0 (Entropy reduction—γ). *In a pair $X^p - X^q$, such X^p and X^q do not have a possible spurious correlation, if $\bar{\gamma}_{pq} > 0$, then orient the edge as: $X^p \to X^q$.*

Furthermore, we make use of the following rule to orient additional edges when the optimal lag $\bar{\gamma}_{pq}$ is null based on the fact that CTMI increases asymmetrically with respect to the increase of λ_{pq} and λ_{qp}. This rule infers the direction of the cause by checking the difference in the window sizes as the window size of the cause cannot be greater than the window size of the effect.

ER-rule 1 (Entropy reduction—λ). *In a pair $X^p - X^q$, such X^p and X^q do not have a possible spurious correlation, if $\bar{\gamma}_{pq} = 0$ and $\bar{\lambda}_{pq} < \bar{\lambda}_{qp}$, then orient the edge as: $X^p \to X^q$.*

We call our method `PCTMI`. The output of the algorithm is a CPDAG version of the summary graph such that all lagged relations are oriented, but instantaneous relations are partially oriented.

PCGCE We make use of the PC algorithm to construct extended summary graphs from observational time series. The skeleton is constructed as in the PC algorithm, using (conditional) GCE to measure the (conditional) dependence between X^p in the past slice to X^q in the present slice, and classical (conditional) mutual information for the edges in the present slice. Orientation is done using PC rules, the overall process being referred as PCGCE. The particularity of this method is that it is the first one to discover an extended summary causal graph from observational time series.

NBCB and CBNB We introduce two classes of algorithms, NBCB and CBNB, designed to infer causal graphs from time series data by combining elements of noise-based and constraint-based approaches. Hybrid frameworks like ours integrate methods from different families to enhance graph inference by overcoming limitations inherent in individual algorithms. We relax the faithfulness assumption required by constraint-based methods by adopting the adjacency faithfulness assumption.

Assumption 3.3.2 (Adjacency Faithfulness, (Ramsey et al., 2006)). *Let $\mathcal{G}^{\mathrm{f}} = (\mathbb{E}^{\mathrm{f}}, \mathbb{V}^{\mathrm{f}})$ be an FTCG. If two nodes X and Y in $\mathbb{V}^{\mathrm{f}}$ are adjacent in $\mathcal{G}^{\mathrm{f}}$, then they are dependent conditionally on any subset of $\mathbb{V}^{\mathrm{f}} \backslash \{X, Y\}$.*

In the NBCB class of algorithms, we first construct a fully connected graph and orient lagged relations based on temporal priority. Then, we determine the causal order among all instantaneous nodes using noise-based methods before testing conditional independences to prune the causal graph.

For the CBNB class of algorithms, we determine the skeleton using conditional independence tests and add temporal orientation. The causal order among instantaneous nodes is established using a clever technique leveraging the knowledge of the skeleton.

Both classes of algorithms yield the true graph under perfect conditional independences. We also explore their robustness against assumption violations.

When working with the summary causal graph, we employ TCE or partial correlation as (conditional) independence tests when assuming linear (conditional) dependencies.

3.3.2 State-of-the-art

Table 3.1 displays the main characteristics of representative algorithms. As one can note, most methods infer a window causal graph. Methods that directly aim at inferring a summary causal graph may have advantage over methods that first infer a window causal graph when considering the summary graph only, being faster and directly aiming at solving a simpler problem. The distinction on the type of graphs inferred is thus not a way to rank causal discovery methods; it just reflects the fact that the objectives differ from one method to the other.

The detection of instantaneous relations is important from a practical point of view as the difference in time between two events associated to two time series may not be observed if the sampling frequencies of the time series are small. Roughly only half of the methods address this particular problem[3]. Being able to detect relations with a gap greater than 1 is also important in practical situations and only oCSE is restricted to a gap of 1. Methods that are not able to infer self causes usually assume that self causes always exist, which seems reasonable in real-life examples.

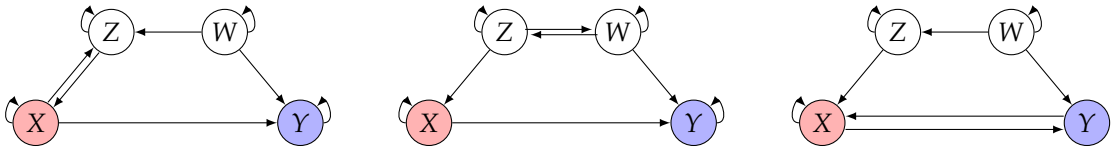
Regarding the type of underlying models, almost all methods rely on a particular model (except constraint-based methods). Among the methods relying on a model, roughly half of them rely on a linear model. Concerning ANLTSM, if the underlying model considered is nonlinear for observed variables, it is linear for hidden ones. Relying on a specific model can be an advantage when the data considered arises from a similar model. It can be of course a disadvantage when this is not the case. Lastly, as one can note, most models use few (less than 5) hyper-parameters, with the exception of TCDF which is based on deep neural networks.

This table serves as a guide to help choose the method that best fits the assumptions underlying the data.

[3]Note that the most recent version of PCMCI includes this possibility. We are discussing here the standard version.

Method	Causal graph	Faithfulness / Minimality	Causal Markov Condition	Instantaneous rel.	Lag > 1	Inference of self causes	Inst. Hidden Conf	Lagged Hidden Conf.	Model based	Linear model	< 5 Hyper-parameters
MVGC (Granger, 1969)	S			✗	✓	✗	✗	✗	✓	✓	✓
TCDF (Nauta et al., 2019)	W			✓	✓	✓	✓	✗	✓	✗	✗
PCMCI (Runge, 2020)	W	F	✓	✗	✓	✓	✗	✗	✗	✗	✓
oCSE (Sun et al., 2015b)	S	F	✓	✗	✗	✓	✗	✗	✗	✗	✓
PCGCE (Assaad et al., 2022a)	E	F	✓	✓	✓	✓	✗	✗	✗	✗	✓
PCTMI Assaad et al. (2022b)	S	F	✓	✓	✓	✓	✗	✗	✗	✗	✓
VarLiNGAM (Hyvärinen et al., 2008)	W	M	✓	✓	✓	✓	✗	✗	✓	✓	✓
TiMINo (Peters et al., 2013)	S	M	✓	✓	✓	✗	✗	✗	✓	✗	✓
DYNOTEARS (Pamfil et al., 2020)	W			✓	✓	✓	✗	✗	✓	✓	✓

Table 3.1: Summary of the main characteristics of representative algorithms in all the families used for illustration in this chapter. For causal graphs, S means that the method provides a summary causal graph, W a window causal graph and E an extended summary causal graph; F corresponds to faithfulness and M to minimality.



(a) $P(y_t|do(x_t))$ (b) $P(y_t|do(x_{t-\gamma}))$ (c) $P(y_t|do(x_{t-1}))$

Figure 3.6: Three SCGs and a total effect which is identifiable. Red and blue vertices in the FTCGs represents the total effect we are interested in.

3.4 Causal reasoning in time series causal graph

We focus in this section on identifying from observational data the total effect of the singleton variable $X_{t-\gamma}$ on the singleton variable Y_t, written $P(Y_t = y_t|do(X_{t-\gamma} = x_{t-\gamma}))$ (as well as $P(y_t|do(x_{t-\gamma}))$ by a slight abuse of notation), when the only knowledge one has of the underlying DSCM consists in the ESCG or SCG derived from the unknown, true FTCG. Y_t corresponds to the response and $do(X_{t-\gamma} = x_{t-\gamma})$ represents an intervention (as defined in (Eichler and Didelez, 2007, Assumption 2.3)) on the variable X at time $t - \gamma$, with $\gamma \geq 0$.

Context Each candidate FTCG proposes a particular decomposition of the true joint probability distribution which is given by the standard recursive decomposition that characterizes Bayesian networks. Not all decompositions are however correct wrt the true probability distribution P. In general, a total effect $P(y_t \mid do(x_{t-\gamma}))$ is said to be identifiable from a graph if it can be uniquely computed with a do-free formula from the observed distribution (Pearl, 1995; Perkovic, 2020). In our context, this means that the same do-free formula should hold in all candidate FTCG so as to guarantee that it holds for the true one.

Definition 3.4.1 (Identifiability of total effects in ESCGs and SCGs)**.** *In a given ESCG or SCG $\mathcal{G}$, $P(y_t \mid do(x_{t-\gamma}))$ is identifiable iff it can be rewritten with a do-free formula that is valid for any FTCG in the set of compatible graphs $\mathcal{C}(\mathcal{G})$.*

One way to rewrite $P(y_t \mid do(x_{t-\gamma}))$ with a do free-formula is by finding an adjustment set of variables for which:

$$P(y_t|do(x_{t-\gamma})) = \sum_{\mathbf{z}} P(y_t|x_{t-\gamma}, \mathbf{z})P(\mathbf{z}). \tag{3.4}$$

Whenever a set of variables satisfy Equation 3.4, we call it a valid adjustment set. The standard backdoor criterion allows one to obtain valid adjustment sets using the true FTCG. We provide here another version of the backdoor criterion that allows us to find a valid adjustment set given all candidate FTCGs without knowing which one is the true FTCG.

Definition 3.4.2 (Backdoor criterion over all candidate FTCGs)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an ESCG or SCG. A set of vertices $\mathcal{Z}$ in $\mathcal{V}$ satisfies the backdoor criterion over all candidate FTCGs relative to $(X_{t-\gamma}, Y_t)$ if*

(i) $\mathcal{Z}$ blocks all backdoor paths between $X_{t-\gamma}$ and Y_t in any FTCG in $\mathcal{C}(\mathcal{G})$,

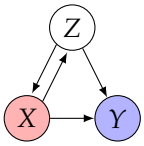
(ii) $\mathcal{Z}$ does not block any directed path between $X_{t-\gamma}$ and Y_t in any FTCG in $\mathcal{C}(\mathcal{G})$,

(iii) $\mathcal{Z}$ does not contain any descendant of X_t in any FTCG in $\mathcal{C}(\mathcal{G})$.

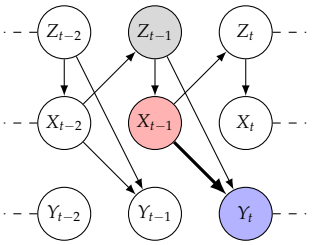
Note that when there is no backdoor path between $X_{t-\gamma}$ and Y_t in any FTCG in $\mathcal{C}(\mathcal{G})$, $\mathcal{Z} = \emptyset$ satisfies the backdoor criterion over all candidate FTCGs.

The backdoor criterion over all candidate FTCGs is sound for the identification of the total effect $P(y_t|do(x_{t-\gamma}))$ in an ESCG or SCG, as stated in the following corollary that can be deduced from Pearl (1995, Theorem 1).

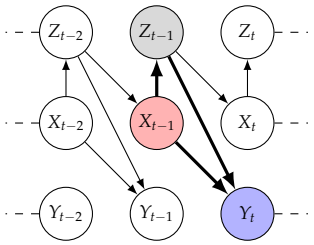
Corollary 1. *Let X and Y be distinct vertices in an ESCG or SCG $\mathcal{G}$ of a DSCM with true (unknown) probability P. Under consistency throughout time for $\mathcal{G}$, if there exists a set $\mathcal{Z}$ satisfying the backdoor criterion over all possible FTCGs relative to $(X_{t-\gamma}, Y_t)$, then the total effect of $X_{t-\gamma}$ on Y_t is identifiable in $\mathcal{G}$, and $\mathcal{Z}$ is a valid adjustment set.*



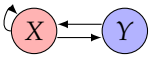
(a) An SCG $\mathcal{G}_1^s$ and the total effect $P(y_t|do(x_{t-1}))$.



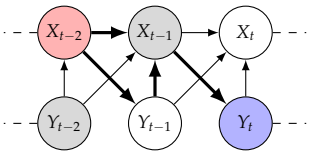
(b) A first candidate FTCG.



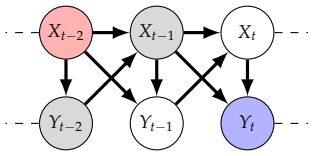
(c) Another candidate FTCG.



(d) An SCG $\mathcal{G}_3^s$ and the total effect $P(y_t|do(x_{t-2}))$.



(e) A first candidate FTCG.



(f) Another candidate FTCG.

Figure 3.7: Two SCGs and a total effect which is not identifiable (on the left). Two candidate FTCGs (middle and right). Each pair of red and blue vertices in the FTCGs represents the total effect we are interested in. Gray vertices are ambiguous: they constitute a backdoor path in the FTCG in the middle and belong to a directed path in the FTCG in the right (bold edges indicate direct paths from $X_{t-\gamma}$ to Y_t).

However, enumerating all candidate FTCGs is computationally expensive (Robinson, 1977), even when considering the constraints given by an ESCG or an SCG.

Identifiability in ESCG The total effect is always identifiable by adjustment in ESCGs.

Theorem 3.4.3. *(Identifiability in ESCG) Consider an ESCG $\mathcal{G}^e$. Under consistency throughout time for $\mathcal{G}^e$, the total effect $P(y_t|do(x_{t-\gamma}))$ is identifiable in $\mathcal{G}^e$ for any $\gamma \geq 0$. Furthermore, the set*

$$\mathcal{B}_\gamma = \{(Z_{t-\gamma-\ell})_{1\leq\ell\leq\gamma_{\max}} | Z_{t^-} \in Par(X_t, \mathcal{G}^e)\} \cup \{Z_{t-\gamma} | Z_t \in Par(X_t, \mathcal{G}^e)\}$$

is a valid adjustment set for $P(y_t|do(x_{t-\gamma}))$.

Identifiability in SCG Here states the main result of this section: sufficient conditions for the identifiability in SCG. Recall that $Cycles(X, \mathcal{G}^s)$ is the set of all directed cycles containing X in $\mathcal{G}^s$, and $Cycles^>(X, \mathcal{G}^s)$ is the subset where cycles contain at least 2 different vertices.

Theorem 3.4.4. *(Identifiability in SCG) Consider an SCG $\mathcal{G}^s = (\mathcal{V}^s, \mathcal{E}^s)$ associated with a DSCM with true (unknown) probability distribution P. Under causal sufficiency and consistency throughout time, the total effect $P(y_t|do(x_{t-\gamma}))$, with $\gamma \geq 0$, is identifiable if $X \notin Anc(Y, \mathcal{G}^s)$ or $X \in Anc(Y, \mathcal{G}^s)$ and one of the following conditions holds:*

1. *$Cycles^>(X, \mathcal{G}^s \backslash \{Y\}) = \emptyset$ and there exists no σ-active backdoor path $\pi^s = \langle V^1 = X, \ldots, V^n = Y\rangle$ from X to Y in $\mathcal{G}^s$ such that $\langle V^2, \ldots, V^{n-1}\rangle \subseteq Desc(X, \mathcal{G}^s)$ or*
2. *$\gamma = 0$ and there exists no σ-active backdoor path $\pi^s = \langle V^1 = X, \ldots, V^n = Y\rangle$ from X to Y in $\mathcal{G}^s$ such that $\langle V^2, \ldots, V^{n-1}\rangle \subseteq Desc(X, \mathcal{G}^s)$ or*
3. *$Cycles^>(X, \mathcal{G}^s \backslash \{Y\}) = \emptyset$ and there exists σ-active backdoor path $\pi^s = \langle V^1 = X, \ldots, V^n = Y\rangle$ from X to Y in $\mathcal{G}^s$ such that $\langle V^2, \ldots, V^{n-1}\rangle \subseteq Desc(X, \mathcal{G}^s)$, and $n = 2$, and $\gamma = 1$, and $Cycles(Y, \mathcal{G}^s \backslash \{X\}) = \emptyset$.*

In the main paper, we have proved the above theorem proving that the following set:

$$\mathcal{A}_\gamma = \{(Z_{t-\gamma-\ell})_{1\leq\ell\leq\gamma_{\max}} | Z \in Desc(X; \mathcal{G}^s)\} \cup \{(Z_{t-\gamma-\ell})_{0\leq\ell\leq\gamma_{\max}} | Z \in \mathcal{V}^s \backslash Desc(X, \mathcal{G}^s)\} \quad (3.5)$$

is a valid adjustment set when the total effect is identifiable. As one can note, it contains all possible parents of $X_{t-\gamma}$ in all candidate FTCGs of $\mathcal{G}^s$. Thus, $\mathcal{A}_\gamma$ blocks any backdoor path π between $X_{t-\gamma}$ and Y_t in any candidate FTCG through the parent of $X_{t-\gamma}$ on that path.

Figure 3.6 show several cases where the total effect is identifiable, and Figure 3.7 show several cases where the total effect is not identifiable.

Chapter 4

Semi-supervised learning

This chapter is the result of collaborations with Massih-Reza Amini (LIG, Computer Science Laboratory) and Valérie Monbet (IRMAR, Mathematics Research Institute of Rennes), and Vasilii Feofanov (PhD student), Lies Hadjadj (PhD student) and Ashna Jose (PhD student). The Python code associated to the methods has been developed by Vasilii Feofanov[a], Lies Hadjadj and Ashna Jose[b]. Thanks to them!

- Self-training: a survey, M.-R. Amini, V. Feofanov, L. Pauletto, L. Hadjadj, E. Devijver and Y. Maximov, (2024+), Neurocomputing, link arXiv
- Classification Tree-based Active Learning: A Wrapper Approach, A. Jose, E. Devijver, M.-R. Amini, N. Jakse, R. Poloni, preprint, link arXiv
- Efficient Initial Data Selection and Labeling for Multi-Class Classification Using Topological Analysis, L. Hadjadj, E. Devijver, R. Molinier, M.-R. Amini, ECAI 2024, link arXiv
- Multi-class probabilistic bounds for self- learning, V. Feofanov, E. Devijver, M.-R. Amini, Journal of Machine Learning Research (2024), link
- Tree-based Quantile Active Learning for automated discovery of MOFs, A. Jose, E. Devijver, R. Poloni, V. Monbet, N. Jakse, AI for Accelerated Materials Design - NeurIPS 2023 Workshop, link
- Regression tree-based active learning, A. Jose, J. Mendonca, E. Devijver, N. Jakse, V. Monbet, R. Poloni (2023). Data Mining and Knowledge Discovery, link
- Wrapper feature selection with partially labeled data, V. Feofanov, E. Devijver, M.-R. Amini (2022). Applied Intelligence, 52(11):12316–12329, link.
- Transductive bounds for the multi-class majority vote classifier, V. Feofanov, E. Devijver, M.-R. Amini (2019). Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):3566–3573, link.

[a] and is available at https://github.com/vfeofanov.
[b] and is available at https://github.com/AshnaJose.

Machine learning algorithms, when dealing with a classification or a regression task, often require large training sets to perform well. However, labeling such large amounts of data is not always feasible, as in many applications, substantial human effort and material cost is needed.

This chapter is focusing on semi-supervised learning (Chapelle et al., 2010), where data are partially-labeled: labels are available only for some of the available training examples. The semi-supervised learning lies inherently between the supervised learning and the unsupervised learning, combining the best of the two worlds in the algorithmic perspective: unlabeled examples contain valuable information about the problem, which improves the performance of the supervised methods, and labeled examples guide the method, which improves the performance and interpretability of unsupervised methods.

Let $\mathbf{X} \in \mathcal{X}^D$ be a D-dimensional vector, and let $Y \in \mathcal{Y}$ be a random variable linked to $\mathbf{X}$. We will consider in this chapter the regression task where $\mathcal{Y} \subset \mathbb{R}$ and the multiclass classification task where $\mathcal{Y} = \{1, \ldots, c\}$. We consider a dataset of size N with all unlabeled observations $\{\mathbf{x}_i\}_{=1}^N$, and we denote $\{y_i\}_{i=1}^N$ the associated (potentially unknown) labels.

The organization of this chapter is the following.

- In Section 4.1, we deal with the active learning task, where one wants to smartly choose the labels to be considered in the training sample. In pool-based active learning, we observe a sample set $X_{\mathcal{U}} = \{\mathbf{x}_i\}_{i=1}^N$ and no label, $Z_{\mathcal{L}} = \emptyset$. We have access to an oracle $\mathcal{O} : \mathcal{X} \to \mathcal{Y}$ that can provide the true label y_i for every observation $\mathbf{x}_i$, for $1 \leq i \leq N$ at some (expensive) cost. We propose two contributions. The first one is for the classification task. Using topological data analysis tools, we construct topological regions, that are homogeneous with the classification. Then, the oracle is labelling few points, that we can distill to provide a (pseudo)-labeled sample. This is a project within the PhD thesis of Lies Hadjadj, supervised by Massih-Reza Amini and Sana Louhichi, and in collaboration with Rémi Molinier. All the details are available in Hadjadj et al. (2024). The second contribution is general, and can deal with classification, regression, and even regression over a region of interest. It is based on standard trees, from which we detect regions to sample. This is the topic of Ashna Jose's PhD thesis, co-supervised with Noel Jakse and Roberta Poloni. All the details are available in Jose et al. (2023).[1]

- In Section 4.2, we consider the multi-class classification task. We assume that we observe few labels, $Z_{\mathcal{L}} = \{\mathbf{x}_i, y_i\}_{i=1}^{\ell}$. We propose a generalization error bound for the majority vote classifier, where unlabeled data are pseudo-labelled (Feofanov et al., 2024), from which we derive a multi-class self-training algorithm[2]. We also deal with high-dimensional data, where the original set of features may contain irrelevant or redundant characteristics to the output, which with the lack of labeled information leads to inefficient learning models. We propose a wrapper feature selection method for a low sparsity level (Feofanov et al., 2022). This section corresponds to the PhD thesis of Vasilii Feofanov, co-supervised with Massih-Reza Amini.

[1] Code available at https://github.com/AshnaJose/Regression-Tree-based-Active-Learning
[2] Code available at https://github.com/vfeofanov/trans-bounds-maj-vote

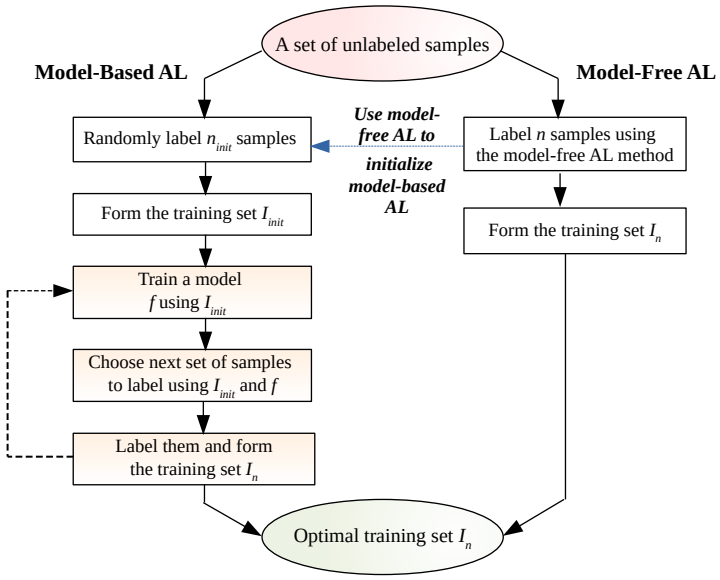


Figure 4.1: Illustration of model-free vs model-based AL methods..

4.1 Active learning

Active learning (AL) aims to detect the observations to be labeled to optimize the learning process and efficiently reduce the labeling cost. The primary assumption behind active learning is that machine learning methods could reach a higher level of performance while using a smaller number of training labels if they were allowed to choose the training dataset (Settles, 2009). Formally, if $\hat{f}_{I_n}$ denotes the estimator among a class $\mathcal{F}$ of prediction models learnt on a training set of size n indexed by $I_n \subset \{1, \ldots, N\}$ with respect to the risk R,

$$\hat{f}_{I_n} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i \in I_n} R(f(\mathbf{x}_i), y_i) \right\}, \tag{4.1}$$

we look for the set of observations indexed by $\mathcal{I}_n \subseteq \{1, \ldots, N\}$ of size n such that the transductive risk is minimized:

$$\mathcal{I}_n = \underset{I_n \subseteq \{1,\ldots,N\}; \# I_n = n}{\operatorname{argmin}} \left\{ \frac{1}{N-n} \sum_{i \notin I_n} R(\hat{f}_{I_n}(\mathbf{x}_i), y_i) \right\}. \tag{4.2}$$

However, in practice, one does not have access to the transductive risk because the labels are not observed.

When no labels are given at first, known as the cold-start problem, only the knowledge of the features can be used. Then, having a first set of labeled data, some methods in AL are proposed to increase the set of labels, that is to find new observations to be labeled to improve even more the models. AL methods can be categorized based on their query criteria into model-free and model-based method O'Neill et al. (2017). The former exploit only the feature space information to construct the most informative training set, while the latter use response information through regression functions trained on previously labeled samples. The global scheme is introduced in Fig. 4.1. The classification task, where $\mathcal{Y} = \{1, \ldots, c\}$, has been more studied in the literature, but few results focus on the regression task. We review in the following the main ideas in the literature, divided into model-free and model-based methods, whatever the nature of the output Y.

Model-free active learning Space filling approaches over the feature space Baram et al. (2004); Wu et al. (2019) come under model-free methods, as they target a training set diverse in features. It selects the sample closest to the centroid of the feature space as the first sample in the training

set, followed by the one farthest from it, according to some distance. The samples to be labeled consequently are the ones farthest from all the samples that have been previously selected to ensure diversity. Representativeness Diversity take into account both diversity and representativeness of the feature space by partitioning the feature space. Several strategy have been proposed: centroids of k-means clustering (Zhu et al., 2008), centroids of agglomerative hierarchical clustering with Ward linkage (Dasgupta and Hsu, 2008), points close to the centroids but satisfying some diversity criterion (Hu et al., 2010; Yu and Hansen, 2017; Liu et al., 2021), and (Zhang et al., 2020) where the closeness is measured using the L_1 distance. In the field of survey methodology, the cube method (Chauvet and Tillé, 2006) has been used for balanced sampling from finite populations. It constructs a sample of fixed size with the same characteristics as those of the features of the full dataset, assuming that this sample will lead to a good approximation of the distribution of the response.

Model-based active learning Model-based AL methods focus on the knowledge of the joint distribution of $\mathbf{X}$ and Y through some (few) labeled examples to mimic the minimization problems given by Eqs. (4.1) and (4.2). To do so, a first set I_{init} of n_{init} samples is detected by a model-free method. Then, $\hat{f}_{I_{\text{init}}}$ is considered as a first estimate of the chosen ML model. This prediction function is now used to select the next $n_{\text{act}} = n - n_{\text{init}}$ samples. Note that the n_{act} samples can be detected sequentially (re-training the model after adding each sample), be split into batches of moderate dimension or be detected in one step. Space filling approach over the feature and response space (Wu et al., 2019) has been proposed, using the Ridge regression to model the predictions. Using the training models, one may want to reduce the expected error: considering the variance reduction (Cohn et al., 1994), reducing the misclassification error (Roy and McCallum, 2001), or reducing the KL divergence (Elreedy et al., 2019) to refer to few of those works. EMCM is looking for samples that will reduce the variance of a considered model (Cai et al., 2013). Query By Committee (QBC) (Riis et al., 2022; McCallum and Nigam, 1998; Yan et al., 2011; Burbidge et al., 2007) is a model-based AL strategy that selects the samples with the highest variance among the predictions from a committee of models. The committee is constructed by bootstrapping on an initial set of passively labeled samples. Mondrian trees is using a purely random tree to model the link between the covariates and the response (Goetz et al., 2018). Some methods have also been proposed based on deep-learning Ren et al. (2021), but this is out of the scope of this work. We mention one that we compare with for the regression task (Holzmüller et al., 2023).

Remark that some of those methods are implemented in Python module (Kottke et al., 2021) for active learning on top of scikit-learn for both the classification and the regression task.

We propose in Section 4.1.1 a method to detect relevant regions to sample, that can be used within already existing AL methods for classification. Then, we propose in Section 4.1.2 a model-based AL method using trees, first for the regression task, that is known to be more difficult (Willett et al., 2005), and then extended to the quantile prediction and to the classification task.

4.1.1 Proper Topological Regions for active learning in classification

We consider a multi-class classification problem such that the input space is $\mathcal{X}^D \subset \mathbb{R}^D$ and the output space is $\mathcal{Y} = \{1, \dots, c\}$ with $c \in \mathbb{N}, c \geq 2$. We assume that close samples (with respect to a distance d) are associated with similar labels, also known as the *smoothness assumption*. In that setting, one can consider neighborhood graphs $(X_{\mathcal{U}}, E)$ on the unlabeled sample $X_{\mathcal{U}} = \{\mathbf{x}_i\}_{i=1}^N$, with E the set of edges. Rips graphs, or more generally Rips complexes (Chazal et al., 2014), can be considered: there is an edge between two vertices if their distance is smaller than some threshold. However, class similarity might be different over the metric space: for example, lower is the density, weaker is the chance to detect a structure within points. Consequently, we introduce Rips graph and σ-Rips graph for an adaptive threshold function σ. Those two notions of neighborhood graph are illustrated in Figure 4.2.

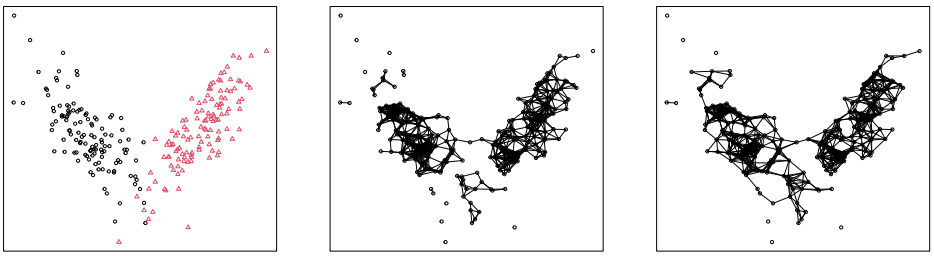


Figure 4.2: (a) A sample of 240 points, generated from a mixture of two bivariate Gaussian distributions. Colors/symboles represent the true classes. (b) Associated Rips graph with $\delta = 0.5$ and d the Euclidean distance. (c) Associated σ-Rips graph, using the parametric form given in Eq. (4.3) with $\delta = 0.5$, $r = 1.08$, $t = 1/5$ and d the Euclidean distance.

Definition 4.1.1 (Rips graph and σ-Rips graph)**.** *Given a finite point cloud $X_{\mathcal{U}} = \{\mathbf{x}_i\}_{i=1}^N$ from a metric space $(\mathcal{X}, d)$ and $\delta \geq 0$, the Rips graph $R_\delta(X_{\mathcal{U}})$ is the graph with set of vertices $X_{\mathcal{U}}$ and whose edges correspond to the pairs of points $(\mathbf{x}_i, \mathbf{x}_j) \in X_{\mathcal{U}}^2$ such that $d(\mathbf{x}_i, \mathbf{x}_j) \leq \delta$.*

Given a real-valued threshold function $\sigma\colon \mathcal{X}^2 \to \mathbb{R}_+^*$, the σ-Rips graph $R_{\sigma(\cdot)}(X_{\mathcal{U}})$ is the graph with set of vertices $X_{\mathcal{U}}$ and whose edges correspond to the pairs of points $(\mathbf{x}_i, \mathbf{x}_j) \in X_{\mathcal{U}}^2$ such that $d(\mathbf{x}_i, \mathbf{x}_j) \leq \sigma(\mathbf{x}_i, \mathbf{x}_j)$.

In this work, we choose the following parametric threshold function:

$$\sigma(\cdot; \delta, r, t)\colon (\mathbf{x}, \mathbf{x}') \in \mathcal{X} \times \mathcal{X} \mapsto \delta(r - \max(\mathbb{P}(\mathbf{x}), \mathbb{P}(\mathbf{x}')))^{\frac{1}{t}} \in \mathbb{R}_+^*, \tag{4.3}$$

with $t \in (0, 1]$ controlling the curvature, $(\delta, r) \in (\mathbb{R}_+^*)^2$ such that $r > \max_{\mathbf{x}} \mathbb{P}(\mathbf{x})$ and are, respectively, dilatation and translation parameters. The max term ensures that σ is symmetric.

In order to detect the underlying topology from a point cloud, our method is based on `ToMATo` (Chazal et al., 2013). This is a clustering method that uses the hill climbing algorithm on the Rips graph along with a merging rule on the Rips graph's persistence. We have adapted their method and extended in Hadjadj et al. (2024) their theoretical results to σ-Rips graph, namely the detection of the peaks for a range of values of τ and their basin of attraction.

The topological regions correspond to the clusters given by `ToMATo` for a σ-Rips graph, defined formally as follows.

Definition 4.1.2. *The topological regions of a sample set $X_{\mathcal{U}}$ coming from an unknown marginal distribution $\mathbb{P}$ and with parameters (a, r, t, τ) are the clusters given by the clustering*

$$\mathrm{TR}_{\delta,r,t,\tau}^{X_{\mathcal{U}},\mathbb{P}} = \texttt{ToMATo}_\tau \left(R_{\sigma(\cdot;\delta,r,t)}(X_{\mathcal{U}}), \mathbb{P} \right).$$

The proper topological regions of a sample set $X_{\mathcal{U}}$ coming from an unknown marginal distribution $\mathbb{P}$ are the topological regions of $\mathrm{TR}_{a^*,r^*,t^*,\tau^*}^{X_{\mathcal{U}},\mathbb{P}}$ where

$$(\delta^*, r^*, t^*, \tau^*) = \underset{(\delta,r,t,\tau)}{\operatorname{argmin}} \left\{ \mathrm{PS} \left(\mathcal{S}, \mathbb{P}, \mathrm{TR}_{\delta,r,t,\tau}^{X_{\mathcal{U}},\mathbb{P}} \right) \right\},$$

with PS *the purity size function, considering the labeling error when propagating the labels inside the topological regions with $\mathcal{L}_{\mathrm{TR}}^{\mathbb{P}}$, penalized by the number of topological regions k in* TR*.*

However, in our AL context, we need to use an unsupervised objective function and we do not want to run `ToMATo` many times for complexity efficiency. We propose in Hadjadj et al. (2024) an algorithm to estimate $(\delta^*, r^*, t^*, \tau^*)$ based on the Silhouette score and the coverage compactness.

We use the proper topological regions in a zero-shot learning algorithm to detect the first examples to be labeled. The strategy is the following: we label the $\mathcal{B}$ largest proper topological

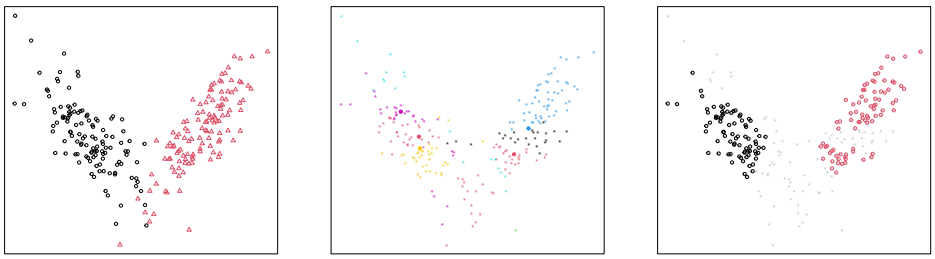


Figure 4.3: Illustration of using PTR in a zero-shot learning algorithm. (a) data with the oracle for the labels. (b) clustering given by tomato, describing the proper topological regions with the best parameters. (c) output of zero-shot learning algorithm, where budget is 5, and with propagation. Not all points are labelled, and not all labels are sure.

regions using label propagation within a cluster: we have to ask to the oracle $\mathcal{B}$ points, and we then label $\sum_{q=1}^{\mathcal{B}} |R_q|$ points. We denote by $\hat{\mathcal{Z}}_{\mathcal{L}}^0$ this first set of labeled points, which includes true labels obtained directly from the oracle, and estimated labels while diffusing the true labels to the topological regions. The benefit to use proper topological regions instead of any clustering method is in details. No structure is assumed, as in k-means for example where clusters have a spherical shape. Here, only the topology is important, thus the algorithm can retrieve connected components even with an ambiguous shape. Moreover, fine hyperparameter tuning in `ToMATo` allows to merge or distinguish between regions. This approach is illustrated in an extensive numerical experimentation in the main paper.

4.1.2 Tree-based Active Learning

Our proposed method relies on standard trees that partition the feature space into hyper-rectangles $(\mathcal{R}_k)_{1\leq k\leq K}$, referred to as regions, and assigns a weight γ_k to each region k:

$$f(\mathbf{x};\Theta) = \sum_{k=1}^{K} \gamma_k \mathbf{1}_{\{\mathbf{x}\in\mathcal{R}_k\}},$$

where $\Theta = ((\mathcal{R}_k, \gamma_k)_{1\leq k\leq K})$. The splitting process, being dyadic, can be represented as a binary tree, where each node determines the features to split on and its corresponding value, resulting in the final partition given by the leaves of the tree.

Active learning for regression

We initialize the method with the indices of the first samples, denoted as I_{init}. A standard regression tree with K leaves is then constructed using the corresponding labeled set $(\mathbf{x}_i, y_i) i \in I_{\text{init}}$, which is subsequently used to predict the response for every unlabeled sample. Following the results derived for purely random Mondrian trees (Goetz et al., 2018), the optimal performance is achieved by selecting n_k^* samples for labeling from each leaf k, where:

$$n_k^* = n_{\text{act}} \frac{\sqrt{\pi_k \hat{\sigma}_k^2}}{\sum_{\ell=1}^{K} \sqrt{\pi_\ell \hat{\sigma}_\ell^2}};$$

where $n_{\text{act}} = n - n_{\text{init}}$, $\hat{\sigma}_k^2$ represents the variance computed on the true labels in leaf k, and π_k denotes the probability that an unlabeled sample $\mathbf{x}_i$ belongs to leaf k. This approach aims to select diverse samples across the response space, ensuring the maximum possible information is captured. The final labeled set is then approximated by $\hat{\mathcal{I}}n = I\text{init} \cup (\cup_{k=1}^{K} I_{\text{act}}^k)$, where I_{act}^k represents the set of samples to be labeled from leaf k.

While random sampling from leaves has been the conventional approach (as done in (Goetz et al., 2018)), it introduces additional randomness and may not fully utilize the labeled samples.

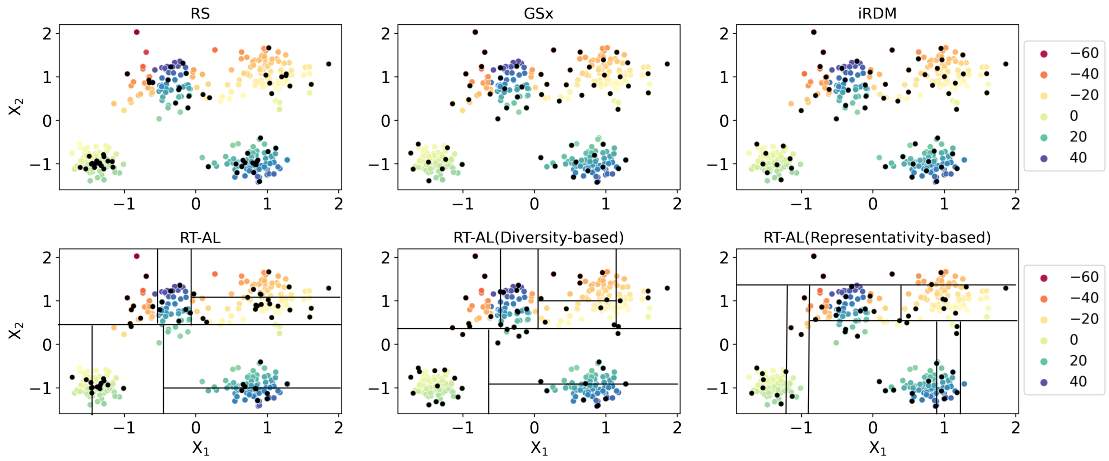


Figure 4.4: Comparison of the samples selected for labeling (shown as black dots) by our method from a generated dataset with 2 features and 500 samples, using different query criteria (labeled as RT-AL, RT-AL(Diversity-based) and RT-AL(Representativity-based)), with passive sampling and model-free AL methods GSx and iRDM. The black lines correspond to the different regions the regression tree splits the feature-response space into. The colors represent the true values of response in the data.

To address this, we propose leveraging ideas from model-free active learning algorithms for sample selection. Points sampled from the leaves can be sampled by feature-space diversity based methods, adapting Wu et al. (2019); or for datasets with prominent clustering, we propose a representativity-based criteria to select the samples from the leaves, adapted from Liu et al. (2021). This is illustrated in Fig. 4.4.

Quantile-based extension

An extension of our method, known as Quantile RT-AL (QRT-AL), focuses on specific quantiles of interest. When subsampling from the leaves, we incorporate the observed response quantiles to oversample the region of interest. The number of samples to be labeled from each leaf k, n_k^*, is distributed considering both the leaf properties and the desired quantile interval:

$$n_k^* = n_{\text{act}} \frac{\sqrt{\pi_k \hat{\sigma}_k^2 \alpha_k}}{\sum_{\ell=1}^{K} \sqrt{\pi_\ell \hat{\sigma}_\ell^2 \alpha_\ell}}, \text{ with } \alpha_k = \frac{\sum_{q=1}^{Q} w^q n_k^q}{\sum_{q=1}^{Q} n_k^q},$$

$(\alpha_k)_{k=1}^K$ specifying the quantile interval of interest, where n_k^q are the number of unlabeled samples in the leaf k in quantile interval q, and w^q are weights defined depending on the quantile of interest.

We demonstrate the performance of our method to predict CO_2 adsorption for MOFs in the hypothetical MOF (hMOF) (Wilmer et al., 2011) database, and to predict band gaps for MOFs in the Quantum MOF (QMOF) (Rosen et al., 2021, 2022) database. These publicly available data sets consist of atomic structures of MOFs along with the respective target properties. We show that this approach decreases the labeling cost tremendously. We also succeed to demonstrate that our approach works for different quantiles of interest, low quantile for band gap predictions and high quantile for predicting adsorption properties.

Extension to classification

We further extend our method to the classification task, where the splitting criterion in the standard trees is based on the entropy, defined by $S_k = -\sum p_{i,k} \log(p_{i,k})$ for the leaf k, with $p_{i,k}$ are the predicted probabilities of each class i in leaf k. The goal is to create leaves as pure as possible, where purity is determined by the distribution of class labels within each leaf.

The method aims to effectively select samples for labeling, prioritizing regions with higher uncertainty or impurity.

We first determine the purity of the leaves. The budget is then divided into the pure leaves and the impure leaves, as a proof of concept in this work we proposed

$$n_{pure} = \frac{n_{act}}{(1 + 3\frac{\max(1, n_{\text{impure leaves}})}{\max(1, n_{\text{pure leaves}})})} \tag{4.4}$$

where n_{pure} is the number of samples to be labeled from the pure leaves, $n_{\text{impure leaves}}$ is the total number of impure leaves and $n_{\text{pure leaves}}$ is the total number of pure leaves. The number of samples to be labeled from each leaf, n_k^*, is computed as:

$$n_k^* = n_{\text{act}} \frac{\sqrt{\pi_k E_k}}{\sum_{\ell=1}^{K} \sqrt{\pi_\ell E_\ell}};$$

where π_k denotes the probability that an unlabeled sample $\mathbf{x}_i$ belongs to leaf k and $E_k = \mathbb{1}_{S_k=0} + \mathbb{1}_{S_k \neq 0} S_k$. Note that using $E_k = 1$ does not imply high entropy for pure regions. Since n_{act} has been distributed among pure and impure regions using Equation (4.4), the criteria to determine n_k^* from the two types of regions are independent: using only the density of unlabeled samples to query new points from pure regions (where entropy is zero), while using both density of unlabeled samples and entropy among the pure labels to query new points from impure regions.

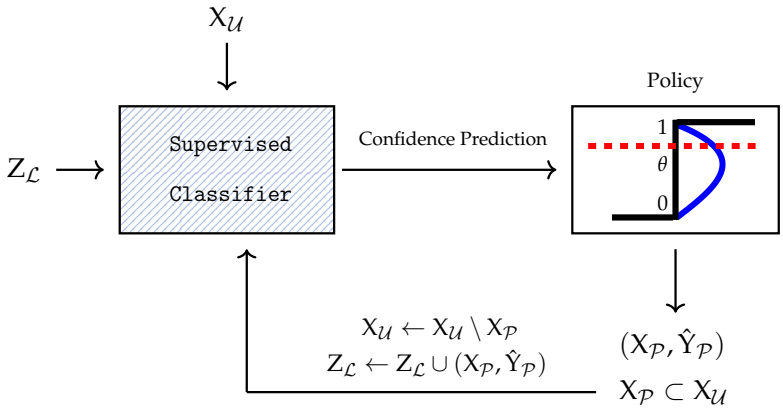


Figure 4.5: Self-training Algorithm: A supervised classifier is learned from the labeled set and is used on the unlabeled set to provide predictions for the classes. A policy determines which predictions are trusted (typically the most confident ones), and pseudo-labels are assigned to those observations based on their predictions. This process is iterated until no more points remain unlabeled.

4.2 Multi-class classification with partially labeled data

In this section, we focus on the multi-class classification task with $\mathcal{Y} = \{1, \ldots, c\}$ with $c \geq 2$. Leveraging a partialy labelled set $Z_{\mathcal{L}} = (\mathbf{x}_i, y_i)_{1 \leq i \leq n}$, our objective is to develop algorithms that effectively utilize both labeled and unlabeled data to improve classification performance.

Self-training, a classic approach in semi-supervised learning (Amini et al., 2024), dates back to the late 1960s (Scudder, 1965; Fralick, 1967). This iterative algorithm, illustrated in Figure 4.5, begins with a supervised classifier trained solely on labeled data. It then iteratively assigns pseudo-labels to unlabeled examples with confidence scores above a predefined threshold, incorporating them into the training set. Recent advancements in self-training include strategies for controlling the number of pseudo-labeled examples through techniques like curriculum learning (Cascante-Bonilla et al., 2021).

Theoretical Foundations Understanding the generalization guarantees of majority vote classifiers is crucial for semi-supervised learning. Many works are focused on deriving tight PAC guarantees for the Gibbs classifier in the inductive case (McAllester, 2003; Maurer, 2004; Catoni, 2007) and in the transductive case (Derbeko et al., 2004; Bégin et al., 2014) for the deterministic case (each unlabeled example is associated to one and only one possible label). While this bound can be tight, it reflects only the individual strength of voters, so using it as a minimization criterion often leads to an increase in the test error (Masegosa et al., 2020). For the majority vote classifier, Amini et al. (2008) derive a transductive bound based on how voters agree on every unlabeled example. Lacasse et al. (2007) propose an upper bound (later called *C-bound*) for the generalization error that is based on the first and the second statistical moments of the margin of the majority vote classifier, as a trade-off between the individual errors of voters (the Gibbs risk) and the error correlation between them. Recently, Frei et al. (2022) derived guarantees for the self-training with a binary linear classifier considering a specific class of mixture models. While extensive research has focused on binary settings, extending these results to the multi-class scenario presents challenges. Existing studies, such as Morvant et al. (2012) and Laviolette et al. (2017), provide insights into generalization bounds and error estimations. However, these studies typically assume perfectly labeled training examples, limiting their applicability to real-world scenarios. While some methods have addressed this issue in both supervised (Natarajan et al., 2013; Scott, 2015a; Xia et al., 2019) and semi-supervised settings (Amini and Gallinari, 2003), theoretical studies in the multi-class case remain limited (Chittineni, 1980).

Feature Selection in Semi-Supervised Learning Given a level of sparsity $d' \ll d$, the goal of feature selection is to find a feature subset $S^* \subseteq \{1, \ldots, d\}$, $|S^*| = d'$ based on $Z_{\mathcal{L}} \cup X_{\mathcal{U}}$ that leads to the highest classification performance among all possible feature subsets of size d'. Most of the semi-supervised feature selection algorithms are extensions of popular supervised or unsupervised filters, scoring features before the construction of a learning model. The Semi-Fisher Score (Yang et al., 2010) extends the supervised Fisher score by embedding the graph Laplacian computed on labeled and unlabeled data. The Semi-supervised Laplacian Score (Zhao et al., 2008) is a graph-based approach that uses unlabeled examples to identify which features are able to preserve the local structure of the data, and labeled examples to maximize distance between observations from different classes. The main disadvantage of filter approaches is that feature importance is evaluated individually, which risks discarding features that are important only in combination with others.(Guyon and Elisseeff, 2003). Recently, embedded approaches, performing model-based feature selection within the training process, are being actively studied. The Rescaled Linear Square Regression (Chen et al., 2017) ranks features by learning a projection matrix using the least square regression with a sparse regularization and scaling the regression coefficients with a set of scale factors. Jiang et al. (2019) use the Bayesian approach to learn weights both for features and unlabeled examples that could be potentially irrelevant. Finally, wrapper approaches use a learner to effectively find a subset of features that are discriminatively powerful together. Ren et al. (2008) proposed a semi-supervised wrapper approach, which incorporates unlabeled data to the training set by means of co-training, and find the best feature subset using forward sequential search. Han et al. (2011) reduced the complexity by pseudo-labeling the unlabeled examples just once and then performing the wrapper feature selection on the augmented the data set. The criterion to detect relevant features is not necessarily limited to the accuracy score, and other learning-based metrics can be used (Song et al., 2007) with both labeled and unlabeled data. However, sequential search approaches (Ren et al., 2008; Han et al., 2011) are time consuming for high-dimensional data. Heuristic search algorithms, like the genetic algorithm (Goldberg and Holland, 1988), significantly reduce the computational time (Siedlecki and Sklansky, 1993; Xue et al., 2015). Syed et al. (2021) proposed genetic algorithm for wrapper selection in the semi-supervised multi-target regression case.

In this section, we aim to contribute to the theoretical understanding and practical application of semi-supervised learning. We introduce in Section 4.2.1 a theoretical study of the majority vote classifier, leading to the development of an adaptive self-training algorithm in Section 4.2.2. Furthermore, we propose a wrapper approach for feature selection in high-dimensional data settings in Section 4.2.3, leveraging the insights gained from our theoretical analysis.

4.2.1 Probabilistic bounds for majority vote classifiers

In this section, we focus on evaluating the error of the majority vote classifier. For a fixed class of classifiers $\mathcal{H} = \{h : \mathcal{X} \to \{1, \ldots, c\}\}$, consider the prior Q_0 and the posterior Q that are defined respectively before and after observing the training set. To measure confidence of the majority vote classifier in its prediction, the notions of class votes and margin are further considered.

$$v_Q(\mathbf{x}, \hat{y}) := \mathbb{E}_{h\sim Q}\mathbb{I}(h(\mathbf{x}) = \hat{y}) = \sum_{h:h(\mathbf{x})=\hat{y}} Q(h);$$
$$m_Q(\mathbf{x}, y) := v_Q(\mathbf{x}, y) - \max_{\hat{y}\neq y} v_Q(\mathbf{x}, \hat{y}).$$

A large value of the vote $v_Q(\mathbf{x}, \hat{y})$ indicates high confidence of the classifier that the true label of $\mathbf{x}$ is $\hat{y}$, while the margin measures the gap between the vote of the true class and the maximal vote among all other classes.

Probabilistic transductive bounds

We define the transductive joint error rate of the Q-weighted majority vote classifier B_Q over the unlabeled set $X_{\mathcal{U}}$ given a vector $\boldsymbol{\theta} = (\theta_y)_{y=1}^{c} \in [0,1]^c$, as:

$$R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q) := \frac{1}{u} \sum_{\mathbf{x}\in X_{\mathcal{U}}} \sum_{\substack{y\in\{1,\dots,c\} \\ y\neq B_Q(\mathbf{x})}} P(Y=y|X=\mathbf{x}) \mathbb{1}_{\{v_Q(\mathbf{x},B_Q(\mathbf{x}))\geq\theta_{B_Q(\mathbf{x})}\}}.$$

In the main paper, we provide a bound on the transductive joint error rate of B_Q by the transductive Gibbs conditional risk, which is used to find adaptive threshold for the self-training algorithm (see Section 4.2.2). We also show that if the Q-weighted majority vote classifier makes most of its error on unlabeled examples with a low prediction vote, under certain conditions, this bound is tight.

Probabilistic C-Bound with Imperfect Labels

When the classifier is trained on both labeled and pseudo-labeled data, there might be some label noise. We consider theoretically a particular scenario when pseudo-labels are random variables $\hat{Y}$, and have been inferred by a teacher model that is trained independently, either by using hold-out set or pre-trained on a similar benchmark. Our framework then could be applied to analyze the first iteration of self-training.

The goal of introducing the random variable $\hat{Y}$ is to understand the difference between the risk of a classifier $h : \mathcal{X} \to \{1,\dots,c\}$, when it is evaluated on the true label Y, $R(h)$, and on the imperfect label $\hat{Y}$, $\hat{R}(h)$. Probabilities $P(\hat{Y}=\hat{y}|Y=y,\mathbf{X}=\mathbf{x})$ are called the mislabeling probabilities, and they allow us to explicitly model imperfection of labels. However, their estimation is very challenging as they depend on $\mathbf{x}$: we assume that the mislabeling probabilities (encoded in the matrix $\mathbf{P} = (p_{\hat{y},y})_{1\leq\hat{y},y\leq K}$) are class-related and instance-independent (Chittineni, 1980; Amini and Gallinari, 2003; Natarajan et al., 2013; Scott, 2015a).

Then, we derive in the main paper a C-bound on the risk of B_Q. Given data with imperfect labels, the direct evaluation of the generalization error rate may be biased, leading to an overly optimistic evaluation. Using the mislabeling matrix $\mathbf{P}$ we derive a more conservative C-bound. In particular, this general result can be used to evaluate the error rate in the semi-supervised setting when mislabeling arises from pseudo-labeling of unlabeled examples. Comparing with the probabilistic transductive bound, the last one directly upper bounds the error rate, so it will be tighter in most of cases.

When the margin mean, the margin variance and the mislabeling matrix are empirically estimated from data, evaluation of the C-bound may be optimistically biased. We analyze the behavior of the estimate with respect to the sample size using the PAC-Bayesian theory initiated by McAllester (1999, 2003). We additionally penalize the C-bound by the sample size and the divergence between Q and Q_0. As u grows, the penalization becomes less severe. The obtained bound may be used to estimate the error of the majority vote from data, with the pseudo-labeled unlabeled examples serving as a hold-out set for estimating the margin moments, and the labeled examples are used to estimate the mislabeling matrix. In the case of classical ensembles, the latter can be performed in the out-of-bag fashion as in (Thiemann et al., 2017; Lorenzen et al., 2019). However, the bound does not appear tighter in practice compared to the supervised case (Laviolette et al., 2017) due to the additional penalization on estimation of the mislabeling matrix.

4.2.2 Multi-class Self-training Algorithm

The central question of applying the self-training algorithm is how to choose the confidence threshold. While setting the threshold to a low value would imply a lot of label noise, setting it to a very high value would put excessive trust in the confidence score initially biased by the small labeled set. Considering the prediction vote of the majority vote classifier as an indicator of confidence, we propose the strategy to automatically select the threshold by minimizing the

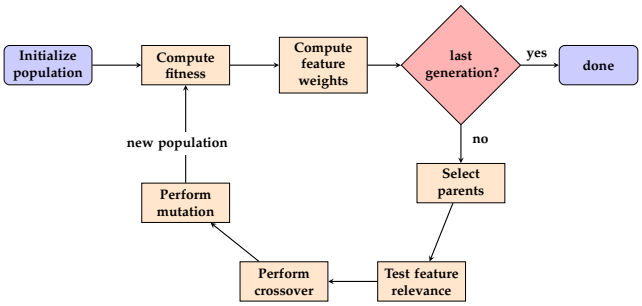


Figure 4.6: Feature Selection Genetic Algorithm (FSGA) in a nutshell.

following criterion $R_{\mathcal{U}|\boldsymbol{\theta}}(B_Q)$ defined as a trade-off between the error we induce by pseudo-labeling and the number of pseudo-labeled examples:

$$R_{\mathcal{U}|\boldsymbol{\theta}}(B_Q) := \frac{R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q)}{\frac{1}{u}\sum_{\mathbf{x}\in \mathrm{X}_{\mathcal{U}}} \mathbf{I}(v_Q(\mathbf{x}, B_Q(\mathbf{x})) \geq \theta_{B_Q(\mathbf{x})}}). \tag{4.5}$$

To evaluate $R_{\mathcal{U}|\boldsymbol{\theta}}(B_Q)$, we bound the numerator of Eq. (4.5) by the transductive bound, where we approximate the posterior $P(Y=y|X=\mathbf{x})$ by $v_Q(\mathbf{x}, y)$ of the base classifier trained on labeled examples only. Although this approximation is optimistic, by formulating the bound as probabilistic we keep some chances for other classes so the error of the supervised classifier can be smoothed. Nevertheless, it must be borne in mind that the hypothesis space should be diverse enough so that the entropy of $(v_Q(\mathbf{x}, y))_{y=1}\{1, \ldots, c\}$ would not be always zero, and the errors are made mostly on low prediction votes. In our experiments, as the base classifier we use the random forest (Breiman, 2001) that aggregates predictions from trees learned on different bootstrap samples.

To find an optimal $\boldsymbol{\theta}^*$ we should perform a grid search over the hypercube $(0, 1]^c$, but we have shown that this K dimensional minimization task can be replaced by K tasks of 1 dimensional minimization, much more cheap.

4.2.3 Wrapper feature selection with partially labeled data

When considering high-dimensional data, we present a semi-supervised framework for wrapper feature selection using both labeled and unlabeled data. The approach consists of two phases: first, we increase variety of the training data by pseudo-labeling the unlabeled examples using a self-training algorithm; then, we perform the feature selection in a wrapper fashion by a proposed genetic algorithm named Feature Selection Genetic Algorithm. Consider the random forest (Breiman, 2001, denoted by RF), due to its ability to output feature weights, versatility for different tasks and high accuracy when the labeled set is scarce (Biau and Scornet, 2016). We use the out-of-bag error, which has been shown to be an unbiased estimator of the generalization error (Breiman, 2001).

In contrast to Ren et al. (2008); Syed et al. (2021), who used a semi-supervised base classifier to evaluate the strength of feature subsets, we assign pseudo-labels to unlabeled examples prior to the feature selection step and then perform a subset search on the expanded training set, which drastically reduce the algorithm's complexity.

After obtaining an augmented data set via TSLA, we perform a heuristic search using a genetic algorithm (Goldberg and Holland, 1988). The classical genetic algorithm, CGA, ignores during the crossover any information like feature importance, since a child inherits features from its parents at random. Moreover, the larger is the number of features, the larger is the search space, so the algorithm becomes highly variable which affects the performance (Xue et al., 2015). In this connection, we propose a new genetic algorithm for feature selection that tackles these two problems: 1) the algorithm takes into account the importance of features during the generation of a new population by using a weighted crossover, 2) it iteratively removes variables that are found to be irrelevant, which accelerates the convergence and reduces the search space. The main steps of this algorithm are summarized in Fig. 4.6.

Chapter 5

Application in Material Science

The line of research developed in this chapter has started with the MAGNET chair of MIAI institute. It is the result of interdisciplinary collaborations with Noël Jakse (SIMAP, Material Science laboratory), Rémi Molinier (Institut Fourier, Mathematics laboratory) and Roberta Poloni (SIMAP), and the supervision of Sébastien Becker (PhD student), Ashna José (PhD student), João Paulo Mendonça (postdoc student) and Johannes Sandberg (PhD student). Thanks to them!

- Feature Selection for High-Dimensional Neural Network Potentials with the Adaptive Group Lasso, J. Sandberg, E. Devijver, N. Jakse, T. Voigtmann, Machine Learning: Science and Technology, 2024, in press, link arXiv
- Informative Training Data for Efficient Property Prediction in Metal-Organic Frameworks by Active Learning, A. Jose, E. Devijver, N. Jakse, R. Poloni, J. Am. Chem. Soc. 146, 6134 (2024), link
- An Artificial Neural Network-based Density Functional Approach for Adiabatic Energy Differences in Transition Metal Complexes, J.P. Almeida de Menonca, A. L. Mariano, E. Devijver, N. Jakse, R. Poloni, J. Chem. Theory Comput. 19, 7555 (2023), link
- Unsupervised topological learning approach of crystal nucleation, S. Becker, E. Devijver, R. Molinier, N. Jakse, Scientific Reports , 2022, link

Materials science is an interdisciplinary field - including chemistry, physics, and computer science - dedicated to the study of the material's properties and performance. Recently, machine learning has found many applications within material science (Schmidt et al., 2019), being used for a variety of tasks such as identification of atomic structures (Becker et al., 2022), prediction of material properties (Furmanchuk et al., 2016; Zheng et al., 2018), among others. We particularly focus on the atomistic scale, to understand how the atoms of the materials are arranged to give rise to molecules, crystal or systems. The chemical bonding and atomic arrangement (crystallography) are fundamental to studying the properties and behavior of any material, and much of the electrical, magnetic and chemical properties of materials arise from this level of structure.

Computational material science is simulating the behavior of materials *in silico* (without experiments), to drastically reduce the time and effort to optimize materials properties for a given application. For example, the crystal nucleation phenomena corresponds to the early stages where the liquid-to-solid transition occurs upon undercooling, initiates at the atomic level on nanometre length and sub-picoseconds time scales and involves complex multidimensional mechanisms with local symmetry breaking that can hardly be observed experimentally in the very details. Unreachable until very recently, experimental observations of early stages of nucleation was achieved by a *tour de force* using time tracking of three-dimensional (3D) Atomic Electron Tomography (Zhou and *et al*, 2019) of metallic nanoparticles. Those complex phenomena remain to date out-of-reach experimentally for bulk systems, thus hindering our theoretical understanding when focusing only on experiments. Two specific methods for computational material science are used here, namely the density functional theory (DFT) and molecular dynamics (MD).

Density-functional theory (DFT), developed in the mid-sixties, has revolutionized the field of materials science. Its integration with advancements in hardware and infrastructure has yielded a powerful tool for predicting the electronic and structural properties of materials, enabling their design and discovery for a wide range of applications (Jones, 2015). Despite the many successes, the implementations of DFT still face several challenges that limit its predictive power and applicability (Cohen et al., 2012). As an example, one can mention the calculation of spin-state energetics in transition metal complexes which represents a challenge for any modern electronic structure *ab initio* method (Wilbraham et al., 2017; Domingo et al., 2010; Radoń, 2019; Swart, 2008; Phung et al., 2018; Reimann and Kaupp, 2023). Recently, several studies have shown how DFT can benefit from machine learning (ML) techniques. Pioneer works (Snyder et al., 2012) proved that it is possible to learn the kinetic energy of 1D fermionic systems. Brockherde et al. (2017) developed a ML scheme to learn the density *via* the external potential/density Hohenberg–Kohn map, so that self consistency can be bypassed. Learning the exchange and correlation functional itself, as demonstrated by the pioneer work by Nagai et al. (2020), has also been studied (Chen et al., 2020; Li et al., 2021; Dick and Fernandez-Serra, 2021; Kirkpatrick et al., 2021). In this class of methods, the exchange and correlation is replaced, adjusted, or corrected by an artificial neural network that receives as input functions of the electronic density. Those computations are still expensive, and one may resort to active learning (introduced in Chapter 4, see Mukherjee et al. (2023); Osaro et al. (2023); Mukherjee et al. (2022); Jablonka et al. (2021)).

An alternative approach is to circumvent DFT by employing classical potentials that offer comparable precision. However, achieving statistically significant results demands large-scale simulations. Despite advancements, conducting million-to-billion-atom simulations for monatomic metals remains challenging, with only a limited number of studies currently providing such datasets (Sosso and *et al*, 2016). For example, when focusing on the crystal nucleation, one should accurately model the interatomic interactions simultaneously in both solid and liquid phases. Classical force fields (Mendelev et al., 2008; Lee et al., 2003) are fast and allow for the study of very large systems containing up to several millions of atoms, but they are often inaccurate and lacking in transferability. In contrast ab initio simulations (Car and Parrinello, 1985), based on DFT (Hafner, 2008), allow for a much more accurate description, and can be applied to any phase of matter and any combination of elements, but at a much higher computational cost, and limited to systems of merely a few hundred atoms. Nucleation events are rare events (viewed from the side of simulations), however, necessitate long simulations of large systems (Sosso and *et al*, 2016). MD is a computer simulation method for analyzing the physical movements of atoms and molecules. The atoms and molecules move as a function of

time due to their mutual interaction (Newton's law), giving a view of the dynamic evolution of the system. MD with generic interaction models (Auer and Frenkel, 2001; ten Wolde et al., 1995).

The organization of this chapter is the following.

- In Section 5.1, a new density functional approach is proposed to predict the adiabatic energy differences in transition metal complexes. Based on an artificial neural network, this method relies on a bio-inspired optimization, due to the non continuous (with respect to parameters) loss function. This corresponds to a part of the postdoc of João Paulo Mendonça, who worked with Noel Jakse, Roberta Poloni and me. More details are available in de Mendonça et al. (2023).

- In Section 5.2, we focus on the construction of a potential for the crystal nucleation. The choice of the descriptors is fundamental to construct the potential, but no consensus has been made to detect a small set of relevant fingerprints. In this work, we propose to include a step of feature selection in the high-dimensional neural network potentials, to let the model select by itself the relevant fingerprints. An Adaptive Group Lasso penalty has been considered. This corresponds to a part of the PhD thesis of Johannes Sandberg, co supervised with Noel Jakse and Thomas Voigtman. More details are available in Sandberg et al. (2022).

- In Section 5.3, we analyze the dynamic of the atoms through the crystal nucleation of metals. Unsupervised learning is used to model the several clusters within the crystal nucleation, allowing to detect early the ones leading to crystal. Topological signatures are used as fingerprints, using persistent homology, to describe local structures required for the clustering methods. This corresponds to a part of the PhD thesis of Sébastien Becker, co supervised with Noel Jakse and Rémi Molinier. More details are available in Becker et al. (2022).

- In Section 5.4, we introduce the result of the active learning method introduced in Section 4.1.2 on two real datasets on Metal-organic frameworks (MOFs) to compute properties (namely band gap and adsorption) that are expensive to obtain. This corresponds to a part of the PhD thesis of Ashna Jose, co supervised with Noel Jakse and Roberta Poloni. More details are available in Jose et al. (2024).

5.1 An Artificial Neural Network-based Density Functional Approach for Adiabatic Energy Differences in Transition Metal Complexes

The aim of the present work is to train an exchange and correlation functional that exhibit high accuracy in the prediction of both electronic densities and adiabatic energy differences of transition metal complexes, as well as atomization energies and densities of simpler molecules. The ML functional is trained using a bio-inspired non gradient-based approach adapted from Particle Swarm Optimization (PSO, Kennedy and Eberhart (1995)). These results show that by training a correction over r2SCAN using three light molecules and three diatomic (metal-nonmetal) transition molecules, the prediction of adiabatic energy differences are improved compared to the state-of-the-art in approximate KS-DFT at no expenses for the performance on atomization energies.

Method Nagai et al. (2020) developed a new functional, defined as a correction over the strongly constrained and appropriately normed (SCAN, Sun et al. (2015a)), trained on energies and densities of a small subset (three molecules only made up by first and second row elements) of the G2 dataset (Curtiss et al., 1997). The result showed an improvement over SCAN in the prediction of ionization potentials and atomization energies on the complete G2 dataset. Yet, previous reports on the literature pointed out some major numerical instabilities in some meta-GGA functionals (Furness and Sun, 2019) including SCAN itself, for electron densities rapidly changing in space. r2SCAN (Furness et al., 2020) recovers most of the physical limits that SCAN originally had. In this work, in order to address the case of transition metal complexes, where rapid fluctuations of the electronic density are expected, a new ANN-based functional defined as a correction over r2SCAN is proposed. We also impose that the exchange and correlation energy density ε_{xc} satisfies the physical limits imposed in r2SCAN using Lagrange multipliers, as done by Nagai et al. (2022). Figure 5.1(left) shows a diagrammatic representation of the feed-forward neural network adopted here. The architecture reflects the choice to adopt a small number of fitted parameters. The local and semilocal functions that are provided to the input neurons (ρ, ζ, s, τ) of the ANN are consistent with a meta-GGA functional. The corrections F_x and F_c are computed separately so that different asymptotic limits are applied to each of them (Sun et al., 2015a; Nagai et al., 2022). For the correct uniform coordinate density-scaling condition (Levy and Perdew, 1985) and the exact spin scaling relation (Oliver and Perdew, 1979) to be satisfied simultaneously, the exchange energy should not depend explicitly in ρ and ζ, so this values are not included as inputs in the calculation of F_x. The loss function is the sum of errors of adiabatic energy differences over three small molecules (NO, H_2O, NH_3), atomization energies over three diatomic transition metal complexes (TMC, here FeO, CuF, CrH), and electron densities (in G2 and TMC). The first term was shown to yield good results for energetics of light element-molecules (Dick and Fernandez-Serra, 2021; Nagai et al., 2020), while training sets with different types of data per sample (i.e., sparse training data) can significantly improve the performance of a machine learned-density functional (Kasim and Vinko, 2021). In order to evaluate the effects of the database, six different training sets were considered: *G2* ([NO + H_2O + NH_3]), *G2+FeO*, *G2+CuF*, *G2+CrH*, *TMC* ([FeO + CuF + CrH]), and *ALL* (*G2 + TMC*). The terms of the loss function only apply when at least one of the related molecules is present in the training set.

Since the equilibrium energies and densities are computed after converging the SCF process with a given set of parameters, the loss function depends on quantities that are related in a non-trivial way (non continuously) to the parameters of the ANN-functional, and smart non-gradient based optimization techniques can outperform gradient-based ones. An adaptation of Particle Swarm Optimization (PSO, Kennedy and Eberhart (1995)) is used as the training algorithm.

Results In Figure 5.2(a), the evolution of the loss function is drawn, where only the best solution (particle) at each migration is reported. After 30 migrations, the solution is converged for every run except for *G2*. For this training set, a monotonic decrease of the loss function is

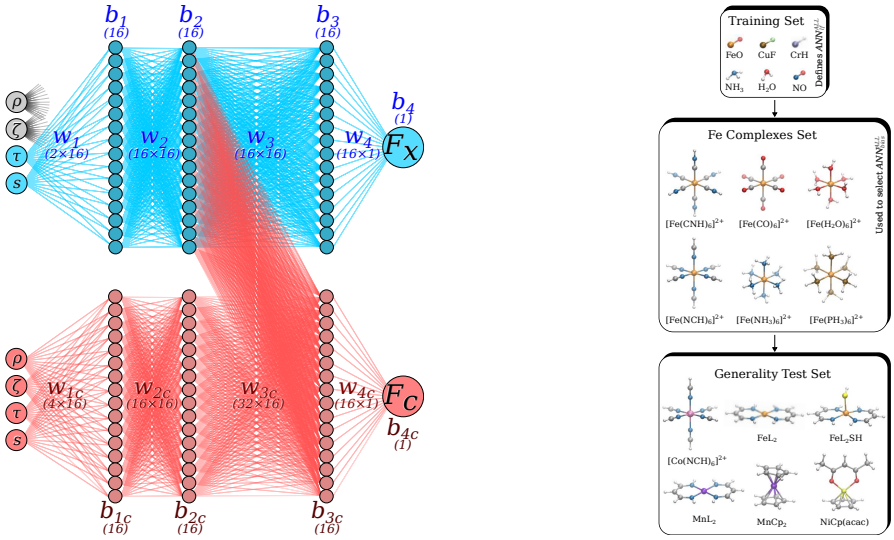


Figure 5.1: Left: Architecture of the artificial neural network used in the present work. The w's and b's represent the weights and bias used on the feed-forward process and the products in parenthesis express the number of parameters in each of those elements. Right: Representation of the different sets of molecular complexes studied in this work.

observed until 36 migrations. A different behavior is observed for the training sets including *TMC*, where the slow monotonic decrease is followed by a sharp decrease, and then by a stationary region. This change in behavior is associated with solutions that cannot be improved *via* migrations.

Figures 5.2(c) and (d) report the heatmap for the Spearman correlation matrix between the partial errors that compose the loss function for particles sampled at migration #15, i.e., during the optimization, and at migration #22, where the optimization stops, for the *ALL* training. At migration #15, in Figure 5.2(c) we see that the errors on energies and densities of FeO and CuF, as well as the densities on the G2 molecules, are optimized simultaneously. The loss function part associated with $\Delta E^{\text{CrH}}_{\text{H-L}}$ and ρ_{CrH} is negatively correlated with those values meaning that their error increase while the total loss function decreases.

In the case of the *ALL* and *TMC* training sets, $\Delta E^{\text{CuF}}_{\text{H-L}}$ was the first to reach a converged value close to zero. At this point of the optimization, the particles dispersion was enough to allow to find a new optimizing path for $\Delta E^{\text{FeO}}_{\text{H-L}}$, as seen in the top left green block in Figure 5.2(d) (positive correlation between these quantities and the total loss function). The second block in Figure 5.2(d), showing strong positive, correlation is negatively correlated with Δ_{err}.

A full analysis of the results is provided in the main paper, comparing this new functional with existed ones. Generally, performance is good, even more on complex molecules, as the square planar compounds MnL_2 and FeL_2, as well as complexes far from the training set in terms of geometry and choice of atoms, such as $[Co(NCH)_6]^{2+}$.

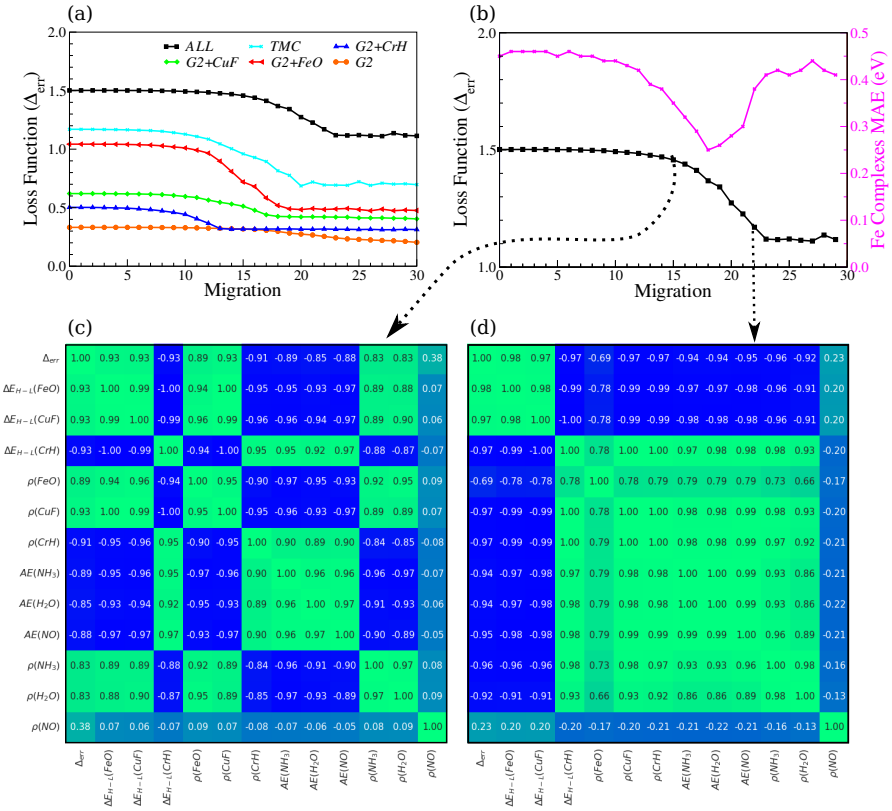


Figure 5.2: (a) Evolution of the loss function values during the training process, for the six different training sets considered here. (b) Comparison between the evolution of the loss function and that of the performance in the Fe complexes test along the migrations in the training with the complete *ALL* training set. (c) and (d) show Spearman's rank correlation coefficients between the total loss function and each one of the L^1 distances that compose it, respectively for the samples visited along migrations #15 and #22.

5.2 Adaptive Selection of Atomic Fingerprints for High-Dimensional Neural Network Potentials

The use of machine-learned interatomic potentials (MLIPs), trained through supervised approaches, allows for bridging the gap between classical force fields and ab initio methods. By training a MLIP on data from DFT-based simulations, it becomes feasible to predict forces with ab initio accuracy while maintaining computational efficiency comparable to classical force fields. This opens up new possibilities to study homogeneous nucleation at the atomic scale.

A fundamental aspect in developing MLIPs within a High-Dimensional Neural Network Potential (HDNNP) framework (Behler and Parrinello, 2007) is the selection of appropriate atomic descriptors (Behler, 2015). These descriptors serve to accurately represent atomic environments by encoding them into a set of features, often referred to as atomic fingerprints. Imbalzano et al. (2018) proposed a feature selection method aimed at optimizing this choice by leveraging filter methods, which do not explicitly incorporate model predictions.

Method The energy of an atom in a material typically depends on its surrounding environment within a few neighbor atom shells, constrained within a cutoff r_c, as depicted in Figure 5.3a. Consequently, it is natural to write the total energy as a sum of local atomic contributions, and a HDNNP is composed of a sum of N NNPs, each associating a local environment with an atomic energy E_i, as illustrated in Figure 5.3b. The atomic NNPs are trained indirectly by fitting the HDNNP to the known total energy derived from DFT-based simulation. Subsequently, differentiation with respect to atomic positions yields force predictions. The inputs of the HDNNP are atomic positions, which are transformed into fingerprint vectors G_i for each

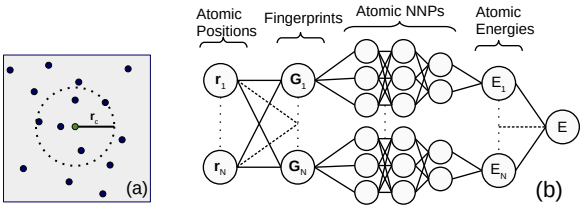


Figure 5.3: The High-Dimensional Neural Network Potential approach. a) Illustration of a local neighborhood of radius r_c around a central atom. b) Illustration of the HDNNP architecture.

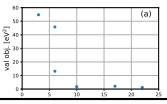
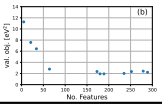
	D	D_{G^5}	MSE (eV^2)	RMSE (meV/atom)	Benchmark (timesteps/s)	Validation objective
S_{HP}	22	10	0.658 ± 0.079	3.16 ± 0.193	0.211	
	10	3	0.702 ± 0.094	3.27 ± 0.223	0.410	
	6	1	1.39 ± 0.25	4.59 ± 0.42	0.581	
S_{large}	64	11	0.250 ± 0.005	1.68 ± 0.02	0.156	
	32	5	0.329 ± 0.033	2.24 ± 0.11	0.267	
	16	2	0.329 ± 0.039	2.23 ± 0.13	0.453	

Figure 5.4: Total number of features (D), number of angular features (D_{G^5}), average test errors with standard deviation, computational performance and validation objective for a sequence of HDNNP models, plotted against the number of selected input features obtained by varying the regularization strength λ. S_{HP}: 22 hand-picked set of features to start with; S_{large}: large set (size: 329) of features to start with.

atom. These vectors are then fed into the atomic NNP to predict atomic energies, which are summed to derive the total energy. In our selection of atomic descriptors, we specifically use the Behler-Parrinello symmetry functions (Behler, 2011), a conventional choice for HDNNPs.

In the realm of linear models, a well-established embedded feature selection method is the Lasso (Tibshirani, 1996), wherein ℓ_1 regularization is applied to the model's input parameters. However, this approach is not directly applicable to Neural Networks. LassoNet (Lemhadri et al., 2021) was recently introduced as an alternative, incorporating bypass connections for each feature and penalizing them by the ℓ_1 norm of the bypass weights. Another option is the Group Lasso (GL), which groups all input weights of each feature and applies regularization based on the Euclidean norm of each such collection of weights.

A refinement to the basic GL approach involves using an adaptive penalty, as proposed by Dinh and Ho (2020). In the adaptive scheme, an initial training is conducted using the standard GL penalty, which provides an initial estimate for the weights. Subsequently, training is repeated using an adaptive penalty. We opt for Adaptive Group Lasso (AGL) over LassoNet, due to its adaptiveness, simpler implementation and fewer hyperparameter.

Results In our experiments, we fix our atomic NNPs to have two hidden layers of 10 nodes each, with tanh activation. Aluminium serves as our illustrative system. We commence with a set of 22 fingerprints, initially chosen manually following the principles outlined in Jakse et al. (2022) and Behler (2015), known to be adequate. Figure 5.4 summarizes the results. It shows the validation objectives plotted against the number of selected features. We observe a plateau, indicating the model's ability to prune redundant features without sacrificing performance, alongside sharp inclines where either essential features are discarded or higher penalties are accepted. From our observations, selecting 10 features proves to be optimal, with 6 also showing potential. These models undergo evaluation on a distinct test set, with ensuing computation of average Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Additionally, we conduct short Molecular Dynamics (MD) simulations of 1000 timesteps for each feature set. Remarkably, we manage to reduce the number of fingerprints without notable accuracy loss in the potential. Furthermore, we notice a preference towards selecting radial (G^2) features,

potentially linked to the simple angular structure of Aluminum.

To evaluate the approach across a broader spectrum, we compile a larger set of 269 radial and 60 angular fingerprints, encompassing the previous 22. While such an extensive fingerprint set may be impractical for HDNNPs due to patently redundant information, Figure 5.4 outlines the results. Although the overall structure remains consistent, we generally select more features with improved performance. Despite the stochastic nature of feature selection, and slight variations in selected features across runs and dataset splits, we find the selection process to be relatively stable.

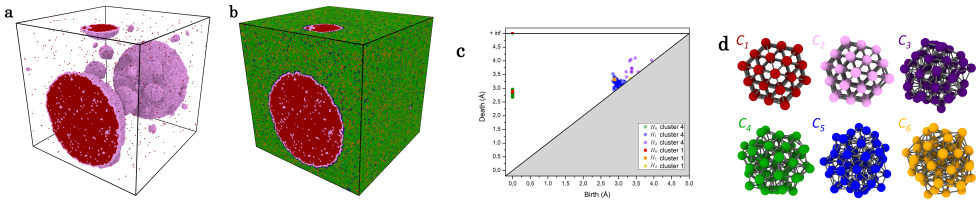


Figure 5.5: Unsupervised learning of homogeneous nucleation. Snapshot of a ten-million atom MD simulation of Ta during nucleation along the $T = 1900$ K isotherm (a and b). (c) Persistence diagrams of homological dimensions H_0, H_1 and H_2 for the mean structures of C_1 (bcc ordering) and C_4 (preponderant liquid structure). (d) Clusters detected by the method. In (a) the snapshot is represented only with atoms in cluster C_1 and cluster C_2 revealing all nuclei, while in (b) atoms of all clusters are displayed: those in cluster C_3 are located mainly at the border of the nuclei and C_4, C_5 and C_6 correspond to the surrounding liquid.

5.3 Unsupervised topological learning for crystal nucleation of Tantalum

Crystal nucleation, the early stages where the liquid-to-solid transition occurs upon undercooling, initiates at the atomic level on nanometre length and sub-picoseconds time scales and involves complex multidimensional mechanisms with local symmetry breaking that can hardly be observed experimentally in the very details. To reveal their structural features in simulations without a priori, an unsupervised learning approach founded on topological descriptors loaned from persistent homology concepts is proposed.

Method Persistent Homology (PH, Carrière et al. (2015); Motta (2018)) is an intrinsically flexible, yet highly informative, tool which detects meaningful topological features deduced from atomic configurations. It was successfully applied to characterise structural environments in metallic glasses (Hirata et al., 2020), ice (Hong and Kim, 2019) and complex molecular liquids (Sasaki et al., 2018). Always used as a structural analysis in these studies, the originality here is to use PH as a translational and rotational invariant descriptor to encode the local structures required for the clustering method. Components of homological dimensions H_0, H_1 and H_2 are then used as the descriptor. The use of two atomic neighbour shells to represent the local environment was shown to optimize the local structural information of descriptors at the expense of a loss of the spatial resolution (e.g., for the averaged bon-orientational order analysis (Lechner and Dellago, 2008)), and here to provide more information in H_0 and H_1 components. Its components are calculated from the Persistent Diagrams (PD) representing the birth and death characteristics of each topological component. The number of H_0 is fixed by the number of neighbour atoms and the number of components of H_1 and H_2 is inferred from a subsampling approach as described in Fasy et al. (2014) to remove the noise. For comparison, the persistent homological informations are depicted on the persistence diagram shown in Figure 5.5(c) for the two mean local structures assigned to C_4 and C_1 depicted in Figure 5.5(d). The differences can be seen here between a disordered liquid structure and a perfect periodic lattice where all the pairs (birth, death) are concurrent for each homological dimensions.

Then, a model-based clustering method is used, namely Gaussian Mixture Models (GMM) (Hastie et al., 2001, Chapter 14) (already used with success to analyse MD simulations (Boattini et al., 2020)) and its estimation by an Expectation Maximization (EM) algorithm (Dempster et al., 1977). The inferred model from the method called hereafter TDA-GMM, is used to identify and describe the structural and morphological properties of the nuclei as well as their liquid environment at various steps of the crystal nucleation.

Results Figure 5.5 depicts the methodology applied to Ta. Crystal nucleation is observed along an isothermal process during which a configuration of the simulation is chosen for the clustering. This configuration contains numerous nuclei of varying sizes and a significant amount of liquid, rendering it representative of the phenomenon. From its inherent structure

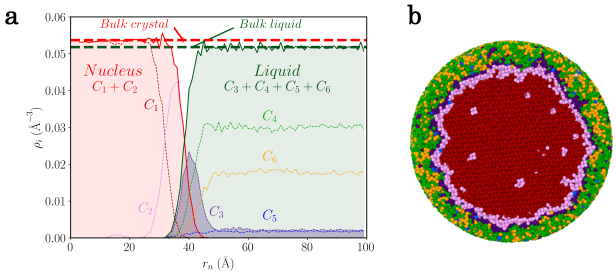


Figure 5.6: (a) Radial density profile of the largest nucleus during the growth at 2.7 nanoseconds along the $T = 1900$ K isotherm. The red (blue) dashed horizontal lines correspond to the average bulk crystalline density (resp. average bulk undercooled liquid without nucleation events), both being simulated at $T = 1900$ K at ambient pressure. (b) Corresponding slice of the nucleus through its centre and the surrounding liquid where atoms have been coloured according to the cluster they belong to.

(Stillinger and T. A. Weber, 1982), a training set of 5,000 non-overlapping local spherical structures within a cutoff radius of 6.8 Å is sampled for unsupervised learning. The sampling is constrained to ensure uniform and random coverage of the entire simulation box. Among all possible sets given by the GMM, the one with 6 clusters shown in Figure 5.5 (d) is representative of the system based on the minimum ICL criterion 5.5(c). The snapshot of the simulation box in Figure 5.5(a) displays only atoms of type C_1 and C_2, as they exhibit clear crystalline order. From this model, each atom of each configuration generated by the MD simulation can be assigned to one of the six clusters (the one with the highest probability), and more than 99.99 % of the structures have a probability to belong to the most probable Gaussian component greater than 0.999.

The nucleation process is characterized by two order parameters: translational order (TO) and bond orientational order (BOO). TO is represented the number density, applied to embryos and nuclei at various growth stages through radial partial atomic density profiles $\rho_i(r)$. Fig. 5.6(a) depicts density profiles $\rho_i(r)$ for all 6 clusters of the largest nucleus and its surrounding liquid at time 2.7 nanoseconds. The corresponding nucleus slice is depicted in Fig. 5.6(b), revealing C_1 atoms at the nucleus center and C_2 atoms at its border. Notably, C_3 atoms are mainly located at the boundary of the nucleus, but they cannot be considered as being part of it, as they are also present in the entire box. The nucleus density matches that of the bulk crystal. Defining the remaining clusters as part of the liquid yields to a total density profile showing minimal influence of the liquid in the nucleus vicinity, maintaining bulk undercooled liquid density.

Each cluster's BOO is identified through Common Neighbor Analysis (CNA, Faken and Jónsson (1994)), revealing C_1 and C_2 clusters with perfect and slightly distorted body-centered cubic (bcc) crystalline ordering. Clusters C_4, C_5, and C_6 exhibit varying degrees of five-fold symmetry (FFS) characteristic of the liquid state, along with non-negligible bcc ordering. Cluster C_3 retains both FFS and bcc order. This BOO of liquid-associated clusters aligns with *ab initio* molecular dynamics simulations (Jakse et al., 2004), interpreted as compatible with the A15 crystalline phase. The CNA-based analysis confirms the TDA-GMM approach's effectiveness in capturing structural intricacies.

5.4 Informative Training Data for Efficient Property Prediction in MOFs

Metal-organic frameworks (MOFs) (Zhou et al., 2012; Wang et al., 2013), formed through coordination bonds between metal ions and organic ligands, are promising materials for efficient gas capture and separation (Ding et al., 2019; Li et al., 1999), due to their ultrahigh porosity, chemical tunability and large surface area (Li et al., 2018, 2009). Recently, they have been shown to be potential candidate materials also for energy storage (Baumann et al., 2019; Zhao et al., 2016; Mariano et al., 2023), water harvesting (Almassad et al., 2022), catalysis (Lee et al., 2009), and sensing (Gamonal et al., 2020), thus evoking an interest in the electronic properties of MOFs (Xie et al., 2020; Johnson et al., 2021; Zhang et al., 2017; Mariano et al., 2023). As such, the identification and/or discovery of novel MOFs with specific properties becomes essential.

To assist in this endeavor, computational techniques such as molecular simulations and density-functional theory were used to screen large MOF datasets, but it is computationally intensive and frequently faces limitations (as illustrated in Section 5.1). Alternatively, machine learning (ML) approaches were exploited to further accelerate MOFs discovery (Demir et al., 2023). Based on a training sample, a descriptor-based ML model is learned, for *e.g.* kernel ridge regression, random forests, or gradient boosting regression trees, (Burner et al., 2020; Janet and Kulik, 2017; Fumanal et al., 2020; Ren and Coudert, 2023; Orhan et al., 2023; Jablonka et al., 2023) to predict properties such as electronic and gas adsorption properties of unseen samples. Recently, deep learning methods such as crystal graph convolutional neural networks (CGCNN, Xie and Grossman (2018); Rosen et al. (2021)) and transformer-based models (Cao et al., 2023; Kang et al., 2023; Park et al., 2023) were also investigated. Despite being powerful and well-suited for large and complex data, deep-learning methods require a substantial amount of labeled data and computational resources to train a complex model. They also require accurate hyperparameter optimizations and sometimes pre-training, which is not feasible when few data are labeled.

In this work, we adopt an opposite strategy to MOFs discovery: we focus on situations where properties are expensive to obtain and therefore large labeled datasets are not available (Nandy et al., 2022). In such cases, it becomes imperative to construct a training set that includes the most diverse, representative, and informative samples. The regression tree-based active learning algorithm introduced in Section 4.1.2 is applied to predict band gap and adsorption properties of MOFs, a novel class of materials that results from the virtually infinite combinations of their building units. Simpler and low dimensional descriptors, such as those based on stoichiometric and geometric properties, found here to better represent MOFs in the low data regime, are used to compute the feature space for this model. The partitions given by a regression tree constructed on the labeled part of the dataset are used to select new samples to be added to the training set, thereby limiting its size while maximizing the prediction quality. Through tests on the QMOF, hMOF, and dMOF data sets, we show that our method constructs small training data sets to learn regression models that predict the target properties more efficiently than existing active learning approaches, and with lower variance. Specifically, our active learning approach is highly beneficial when labels are unevenly distributed in the descriptor space and when the label distribution is imbalanced, which is often the case for real world data. This offers a unique tool to efficiently analyze complex structure-property relationships in materials and accelerate materials discovery.

We report here the results on the QMOF dataset, more details are given in the main paper. The performance of our active learning method, RT-AL, is assessed using the ST-120 descriptor. The MAE for band gap predictions on the held out test set as a function of training set size is reported in Figure 5.7 (a) for RT-AL and other active learning approaches. RT-AL is the best performer for all training set sizes. To further understand the reason behind the good performance of the RT-AL method, we compute an unsupervised structural dimensionality reduction performed using the Uniform Manifold Approximation and Projection (UMAP, McInnes et al. (2018)), with a distance matrix obtained from the ST-120 feature-set of the QMOF data set. The result is reported in Figure 5.7(b) and (c) and the colors on the UMAP represent the values of band gaps and the leaves, respectively. In these figures, we also show 60 labeled samples selected using RT-AL as black circles. Although some clusters in the UMAP space are carried

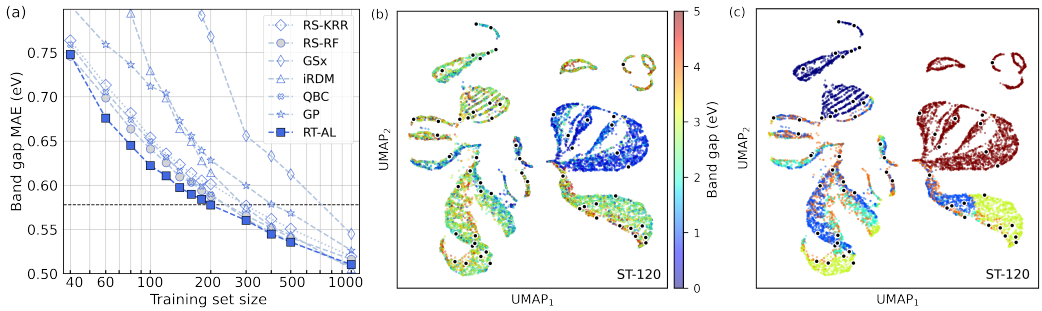


Figure 5.7: (a) MAE for predicting band gaps on the test set as a function of training set size for ST-120 descriptor, using random sampling with KRR (RS-KRR) and RF (RS-RF), and the active learning methods GSx, iRDM, QBC, GP and RT-AL with RF. Each point is an average over 40 runs with different seeds for the train-test split. The horizontal dotted line is a guide to the eye to compare the reduction in labeling cost for RT-AL over other sampling methods. (b) Dimensionality reduction of the training pool of the QMOF data set performed using UMAP, with a distance matrix obtained from the ST-120 feature-set of MOFs in the data set. Colors represent the values of band gap and (c) the different regions given by the regression tree.

forward to the target space, some others have a hint of all colors. This implies that the data is not well clustered in the target space, and neither evenly distributed. RT-AL uses both the input and the target information through the structure of the regression tree and thus it selects MOFs from every region of the target space, and is eventually able to give better predictions for all band gap values. Importantly, RT-AL ensures to sample from all regions of the target space also for very small training sets. In addition, Figure 5.7 (c) shows that the samples selected by RT-AL are distributed well in the feature space, as well as among the various regions (leaves). The tree succeeds to find meaningful patterns (chemical similarity) in the data as shown by how the leaves are distributed in the UMAP.

Conclusion and perspectives

This chapter concludes this manuscript and outlines my future projects and overall view of research. Instead of delving into detailed extensions of the introduced methods, as these perspectives are already covered in the corresponding papers, I will discuss the main projects I am currently involved in and the directions that interest me as I write this thesis.

Functional data analysis

In Chapter 1, we proposed several models for functional data analysis. Here, I highlight some open questions that I intend to pursue in the (near) future.

Fixed discrete time points In all the models we proposed, the time points have been fixed (non-random). However, it would be interesting to understand how the methodology could be generalized to random time points. This question arose during my postdoc, introduced in Section 1.2.1, and remains of interest. Typically, one would need to assume that the distribution of the random time points is regular enough (i.e., no empty regions in the observed pattern of discretized time points). This generalization, though mainly theoretical, is very useful to understand the link with dynamical system.

Oracle inequality for model selection for confidence bands Section 1.2.3 discussed simultaneous confidence bands for linear models and the challenges of dealing with biased estimator. We proposed a heuristic model selection criterion that balances bias and variance to select the dimension in an orthonormal functional basis. This criterion is totally heuristic, and I would like to go deeper in its theoretical analysis, by providing an oracle inequality to ensure the good behavior of the criterion. I aim to deepen its theoretical analysis by providing an oracle inequality to ensure the criterion's effectiveness. The main challenge lies in the supremum norm used to ensure the confidence band's level over the entire time. Classical results are stated for the L_2 norm, so new tools are required to address this.

Causality

In Chapter 3, we addressed several problems related to discovering causal graphs for time series and reasoning about them. This opens many new avenues for research, some of which I will describe here.
Several projects have already begun, supported by two recent grants: my MIAI chair on causality and the Causalit-AI project in the PEPR IA, for which I am the local PI.
I would also like to mention that most of this manuscript has been written during a research stay at the Copenhagen Causality Lab (CoCaLa), which emphasize that my main research direction in the near future will be about causality.

Difference graphs Through discussions with experts, particularly in healthcare, I have learned that difference graphs are of great interest but are not well studied methodologically. This is the focus of Daria Bystrova's postdoctoral work. When observing two populations (e.g., healthy versus sick individuals), experts are particularly interested in the difference between these populations. Usually, they compare the two populations using statistical tests for example. In terms of causal discovery, the naive approach is to construct a causal graph for each group and manually compare the differences. However, it is more efficient to directly infer the difference graph, which is expected to be sparse if the two populations exhibit similar behavior. This approach leads to different assumptions and modeling challenges, such as moving away from the classical faithfulness assumption. The resulting graph is also not causal, but it would be interesting to come back to the structural causal model to understand the difference.

Abstract graphs The identifiability results introduced in Section 3.4 can be seen as a continuity of Perkovic (2020), who discuss identifiability in the Markov Equivalence Class. The graphs we consider (candidate FTCGs) may have different skeletons and may not all be compatible with the true underlying distribution for a given ESCG or SCG, and our results are more general in that sense. Additionally, in ESCGs and SCGs, each vertex does not necessarily correspond to a single observed variable. This can be viewed as an abstract graph where some information is missing (specifically, the lag between the cause and the effect), resulting in several causal graphs (FTCGs) that correspond to different potential skeletons and orientations. I aim to pursue this line of research by considering generic abstract graphs where some information is missing, but we have some knowledge on the causal graph. Anand et al. (2023) explored this for cluster DAGs, assuming acyclicity, which I think is too strong for abstraction. Clément Yvernes is currently doing an internship on this topic and will start a PhD in September 2024 with Marianne Clausel, Eric Gaussier and myself.

Markov blanket Causal discovery can be seen as a way to identify relevant features for a specific node, related to feature selection via the Markov blanket (the set comprising its parents, children and spouses). Théotime Le Goff recently began a PhD on this topic. Several papers in the literature have bridged these two categories. Rütimann and Bühlmann (2009) proposed a graphical model inference based on PC algorithm, initially designed for causal discovery. Causal inference has been proposed based on invariance in prediction in Peters et al. (2016). We aim to relate tools like causal random forest (Wager and Athey, 2018) or BART (Hahn et al., 2020) to feature selection, thereby improving the general model.

Dynamic system and causal inference During my stay at the Copenhagen Causality Lab (CoCaLa), I gained new insights into linking causal discovery for time series and dynamic systems in continuous time. The main drawback of modeling a stochastic process through time series is the sensitivity of the graphical model to the choice of timestamps, the interval between the measurements (Didelez, 2003). Larger intervals correspond to marginalizing over the time in between, creating additional correlations due to common causes or mediating events, but also hiding genuine short-term correlations. Instantaneous causality can also result from meantime effects or unobserved processes. Therefore, one may consider generalizing dynamic dependencies to the continuous-time situation. Formalizing a causal model for dynamical systems, however, is a complex question (Peters et al., 2022). For example, in the iid case, conditional

independence relates to the properties of the graph through Markov condition, but it is not clear how this applies in dynamical models. Local independence has been used for specific model classes (Aalen, 1987; Mogensen and Hansen, 2022). I aim to explore how discrete-time and continuous-time relate and their usefulness in different scenarios. Additionally, delving deeper into SDE and stochastic process theory is an exciting avenue, allowing me to continue developing models for functional data analysis and time series.

Semi-supervised learning

In Chapter 4, we explored several aspects of the semi-supervised learning paradigm. Here, I present several open questions that interest me

Theoretical guarantees for active learning We proposed two methods to construct the training set, suitable for classification, regression or specific value range regression. These works highlight the need of theoretical guarantees for active learning methods. Generalization bounds would ensure good performance, but few results exist in the literature. We also hope for a faster convergence rate than passive learning, as predicted by the central limit theorem, though this remains an open question. On the methodological part, the regression task is rarely studied in the literature, despite its practical importance. Our initial work has shown promising benefits, but the theoretical and practical limitations need further study.

Active learning and semi-supervised learning We focused on semi-supervised learning for multi-class classification. We derived several theoretical bounds, controlling the probabilistic transductive risk and taking into account potential noisy labels, as well as PAC-bayesian bounds on the C-bound. Based on this result, we proposed a multi-class self-training algorithm where the threshold for selecting unlabeled data to pseudo-label is automatically determined.

Semi-supervised learning is taking the benefit of labeled examples as well as the knowledge of unlabeled examples. Pseudo-labeling unlabeled data (with some confidence) can enhance the performance of supervised method, but it is crucial to manage the induced noise carefully.

One perspective is to combine active learning and semi-supervised learning, from both methodological and theoretical viewpoints, to handle real dataset that are expensive to label, as it is the case, for example, in materials science.

Material Science

We began Chapter 5 by highlighting that machine learning has found numerous applications within material science, being utilized for tasks ranging from the fundamental quantum description of matter to the discovery of materials with desired properties. Our findings suggest that machine learning can not only uncover new applications within material science but also drive the development of new methodologies, as discussed in Section 4.1.2.

This interdisciplinary research between machine learning and materials science has yielded significant contributions, leading to both methodological advancements in machine learning and tangible progress in materials science. Specific questions have emerged from each problem, alongside more general inquiries.

One future direction is the Ph.D. project of Vsevolod Morozov, who started his PhD in May 2024 and is supervised by Noel Jakse, Charlotte Laclau and myself. Our goal is to refine the method proposed in Section 5.3 to discern the distinct stages of crystallization in a metal. This entails developing a clustering method with graph embeddings to describe the local neighborhood of each atom while considering the time evolution in the clustering process. This endeavor establishes connections with network inference outlined in Chapter 2 and the link between dynamic system and causality described previsouly.

More broadly, several open questions emerge: How can we integrate experimental results to improve or guide machine learning? How can we leverage insights from physical and chemical domains to enhance machine learning methods?

Bibliography

Aalen, O. O. (1987). Dynamic modelling and causality. Scandinavian Actuarial Journal, 1987(3-4):177–190.

Ahmad, A. and Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. Ieee Access, 7:31883–31902.

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723.

Alkhoury, S., Clausel, M., Devijver, E., Gaussier, É., and Seiller, A. (2024). Ensembles of Probabilistic Regression Trees. working paper or preprint.

Alkhoury, S., Devijver, E., Clausel, M., Tami, M., Gaussier, E., and Oppenheim, g. (2020). Smooth and consistent probabilistic regression trees. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, Advances in Neural Information Processing Systems, volume 33, pages 11345–11355. Curran Associates, Inc.

Allen, G. I. and Liu, Z. (2013). A local Poisson graphical model for inferring networks from sequencing data. IEEE Transactions on NanoBioscience, 12(3):189–198.

Almassad, H. A., Abaza, R. I., Siwwan, L., Al-Maythalony, B., and Cordova, K. E. (2022). Environmentally adaptive MOF-based device enables continuous self-optimizing atmospheric water harvesting. Nat. Commun., 13(1):4873.

Amblard, P.-O. and Michel, O. J. J. (2013). The relation between granger causality and directed information theory: A review. Entropy, 15(1):113–143.

Ambroise, C., Chiquet, J., and Matias, C. (2009). Inferring sparse Gaussian graphical models with latent structure. Electronic Journal of Statistics, 3:205–238.

Amini, M., Laviolette, F., and Usunier, N. (2008). A transductive bound for the voted classifier with an application to semi-supervised learning. In Advances in Neural Information Processing Systems, pages 65–72.

Amini, M.-R., Feofanov, V., Pauletto, L., Hadjadj, L., Émilie Devijver, and Maximov, Y. (2024). Self-training: A survey. Neurocomputing, page 128904.

Amini, M.-R. and Gallinari, P. (2003). Semi-supervised learning with explicit misclassification modeling. In Gottlob, G. and Walsh, T., editors, IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003, pages 555–560. Morgan Kaufmann.

Anand, T. V., Ribeiro, A. H., Tian, J., and Bareinboim, E. (2023). Causal effect identification in cluster dags. Proceedings of the AAAI Conference on Artificial Intelligence, 37(10):12172–12179.

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). A characterization of markov equivalence classes for acyclic digraphs. Annals of Statistics, 25(2):505–541.

Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. Journal de la Société Française de Statistique, 160(3).

Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. Journal of Machine Learning Research, 10:245–279.

Assaad, C. K., Devijver, E., and Gaussier, E. (2022a). Discovery of extended summary graphs in time series. In Cussens, J. and Zhang, K., editors, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, volume 180 of Proceedings of Machine Learning Research, pages 96–106. PMLR.

Assaad, C. K., Devijver, E., and Gaussier, E. (2022b). Entropy-based discovery of summary causal graphs in time series. Entropy, 24(8).

Assaad, C. K., Devijver, E., and Gaussier, E. (2022c). Survey and evaluation of causal discovery methods for time series. J. Artif. Int. Res., 73.

Assaad, C. K., Devijver, E., Gaussier, E., and Ait-Bachir, A. (2021). A mixed noise and constraint-based approach to causal inference in time series. In Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J., and Lozano, J. A., editors, Machine Learning and Knowledge Discovery in Databases. Research Track, pages 453–468, Cham. Springer International Publishing.

Assaad, C. K., Devijver, E., Gaussier, E., Goessler, G., and Meynaoui, A. (2024). Identifiability of total effects from abstractions of time series causal graphs. In The 40th Conference on Uncertainty in Artificial Intelligence.

Auer, S. and Frenkel, D. (2001). Prediction of absolute crystal-nucleation rate in hard-sphere colloids. Nature, 409:1020–1023.

Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008, pages 33–40.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. Journal of Machine Learning Research, 9:485–516.

Bar-Hen, A. and Poggi, J. M. (2016). Influence measures and stability for graphical models. Journal of Multivariate Analysis, 147:145–154.

Barabási, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature Reviews Geneticss, 12(1):56–68.

Baram, Y., El-Yaniv, R., and Luz, K. (2004). Online choice of active learning algorithms. J. Mach. Learn. Res., 5:255–291.

Baraud, Y., Giraud, C., and Huet, S. (2009). Gaussian model selection with an unknown variance. The Annals of Statistics, 37(2):630–672.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. Statistics and Computing, 22(2):455–470.

Baumann, A. E., Burns, D. A., Liu, B., and Thoi, V. S. (2019). Metal-organic framework functionalization and design strategies for advanced electrochemical energy storage devices. Commun. Chem., 2(1).

Becker, S., Devijver, E., Molinier, R., and Jakse, N. (2022). Unsupervised topological learning approach of crystal nucleation. Sci. Rep, 12(3195).

Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F. (2014). PAC-Bayesian Theory for Transductive Learning. In Kaski, S. and Corander, J., editors, Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, volume 33 of Proceedings of Machine Learning Research, pages 105–113, Reykjavik, Iceland. PMLR.

Behler, J. (2011). Atom-centered symmetry functions for constructing high-dimensional neural network potentials. J. Chem. Phys, 134.

Behler, J. (2015). Constructing high-dimensional neural network potentials: A tutorial review. Int. J. Quantum Chem, 115.

Behler, J. and Parrinello, M. (2007). Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys. Rev. Lett, 98(146401).

Beirlant, J., Dudewicz, E. J., Györfi, L., Van der Meulen, E. C., et al. (1997). Nonparametric entropy estimation: An overview. International Journal of Mathematical and Statistical Sciences, 6(1):17–39.

Bellman, R. and Kalaba, R. (1957). Dynamic programming and statistical communication theory. Proceedings of the National Academy of Sciences of the United States of America, 43(8):749.

Berrett, T. B. and Samworth, R. J. (2019). Nonparametric independence testing via mutual information. Biometrika, 106(3):547–566.

Berrett, T. B., Samworth, R. J., and Yuan, M. (2019). Efficient multivariate entropy estimation via k-nearest neighbour distances. The Annals of Statistics, 47(1):288 – 318.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(1):175–197.

Berry, K. J., Johnston, J. E., and Mielke, P. W. (2018). Permutation statistical methods. In The Measurement of Association, pages 19–71. Springer.

Biau, G. and Scornet, E. (2016). A random forest guided tour. Test, 25(2):197–227.

Bickel, P. and Levina, E. (2008). Regularized estimation of large covariance matrices. The Annals of Statistics, 36(1):199–227.

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.

Bigot, J. (2013). Fréchet means of curves for signal averaging and application to ECG data analysis. Ann. Appl. Stat., 7(4):2384–2401.

Birgé, L. (2005). A new lower bound for multiple hypothesis testing. IEEE Transactions Information Theory, 51(4):1611–1615.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. Probability Theory & Related Fields, 138(1-2).

Birgé, L. and Massart, P. (2001). Gaussian model selection. Journal of the European Mathematical Society, 3(3):203–268.

Blein-Nicolas, M., Devijver, E., Gallopin, M., and Perthame, E. (2024). Nonlinear network-based quantitative trait prediction from biological data. Journal of the Royal Statistical Society Series C: Applied Statistics, page qlae012.

Blein-Nicolas, M., Negro, S. S., Balliau, T., Welcker, C., Cabrera-Bosquet, L., Nicolas, S. D., Charcosset, A., and Zivy, M. (2020). A systems genetics approach reveals environment-dependent associations between SNPs, protein coexpression, and drought-related traits in maize. Genome Research, 30(11):1593–1604.

Boattini, E. et al. (2020). Autonomously revealing hidden local structures in supercooled liquids. Nat. Commun., 11:1–9.

Bodinier, B., Filippi, S., Nøst, T. H., Chiquet, J., and Chadeau-Hyam, M. (2023). Automated calibration for stability selection in penalised regression and graphical models. Journal of the Royal Statistical Society Series C: Applied Statistics, page qlad058.

Bouveyron, C., Bozzi, L., Jacques, J., and Jollois, F.-X. (2017). The Functional Latent Block Model for the Co-Clustering of Electricity Consumption Curves. Journal of the Royal Statistical Society: Series C Applied Statistics.

Bouveyron, C., Casa, A., Erosheva, E., and Menardi, G. (2021a). Co-clustering of Time-Dependent Data via the Shape Invariant Model. Journal of Classification.

Bouveyron, C., Côme, E., and Jacques, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. The Annals of Applied Statistics, 9(4):1726–1760.

Bouveyron, C. and Jacques, J. (2011). Model-based Clustering of Time Series in Group-specific Functional Subspaces. Advances in Data Analysis and Classification, pages 281–300.

Bouveyron, C., Jacques, J., Schmutz, A., Simoes, F., and Bottini, S. (2021b). Co-Clustering of Multivariate Functional Data for the Analysis of Air Pollution in the South of France. Annals of Applied Statistics.

Brault, V., Devijver, É., and Laclau, C. (2024). Mixture of segmentation for heterogeneous functional data. Electronic Journal of Statistics, 18(2):3729 – 3773.

Breiman, L. (2001). Random forests. Machine Learning, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). Classification and Regression Trees. Chapman & Hall, New York.

Brockherde, F., Vogt, L., Li, L., Tuckerman, M. E., Burke, K., and Müller, K.-R. (2017). Bypassing the kohn-sham equations with machine learning. Nat. Commun., 8(1):872.

Buhlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bunea, F., Ivanescu, A. E., and Wegkamp, M. H. (2011). Adaptive inference for the mean of a gaussian process in functional data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4):531–558.

Burbidge, R., Rowland, J. J., and King, R. D. (2007). Active learning for regression based on query by committee. In Yin, H., Tino, P., Corchado, E., Byrne, W., and Yao, X., editors, Intelligent Data Engineering and Automated Learning - IDEAL 2007, pages 209–218, Berlin, Heidelberg. Springer Berlin Heidelberg.

Burner, J., Schwiedrzik, L., Krykunov, M., Luo, J., Boyd, P. G., and Woo, T. K. (2020). High-performing deep learning regression models for predicting low-pressure co2 adsorption properties of metal–organic frameworks. The Journal of Physical Chemistry C, 124(51):27996–28005.

Cabeli, V., Verny, L., Sella, N., Uguzzoni, G., Verny, M., and Isambert, H. (2020). Learning clinical networks from medical records based on information estimates in mixed-type data. PLOS Computational Biology, 16(5):1–19.

Cai, T., Zhang, C.-H., and Zhou, H. (2010). Optimal rates of convergence for covariance matrix estimation. The Annals of Statistics, 38(4):2118–2144.

Cai, W., Zhang, Y., and Zhou, J. (2013). Maximizing expected model change for active learning in regression. In 2013 IEEE 13th International Conference on Data Mining, pages 51–60.

Cambanis, S., Huang, S., and Simons, G. (1981). On the theory of elliptically contoured distributions. Journal of Multivariate Analysis, 11(3):368 – 385.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. Ann. Statist., 35(6):2313–2351.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2009). Robust principal component analysis? Journal of ACM, 58(1):1–37.

Cao, Z., Magar, R., Wang, Y., and Barati Farimani, A. (2023). Moformer: Self-supervised transformer model for metal–organic framework property prediction. Journal of the American Chemical Society, 145(5):2958–2967.

Car, R. and Parrinello, M. (1985). Unified approach for molecular dynamics and density-functional theory. Phys. Rev. Lett, 55:2471.

Carlsson, G. and Mémoli, F. (2010). Characterization, stability and convergence of hierarchical clustering methods. Journal of Machine Learning Research, 11:1425–1470.

Carrière, M., Oudot, S. Y., and Ovsjanikov, M. (2015). Stable topological signatures for points on 3d shapes. Eurographics Symp. Geom. Process, 34:1–12.

Carroll, C., Müller, H.-G., and Kneip, A. (2020). Cross-component registration for multivariate functional data, with application to growth curves. Biometrics, to appear.

Cascante-Bonilla, P., Tan, F., Qi, Y., and Ordonez, V. (2021). Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 6912–6920.

Catoni, O. (2007). Pac-bayesian supervised classification: the thermodynamics of statistical learning. arXiv preprint arXiv:0712.0248.

Chakraborty, A. and Panaretos, V. M. (2021). Functional registration and local variations: identifiability, rank, and tuning. Bernoulli, 27(2):1103–1130.

Chamroukhi, F. (2016). Piecewise Regression Mixture for Simultaneous Functional Data Clustering and Optimal Segmentation. Journal of Classification, 33(3):374–411.

Chandrasekaran, V., Sanghavi, S., Parrilo, P., and Willsky, A. (2011). Rank-sparsity incoherence for matrix decomposition. SIAM Journal on Optimization, 21(2):572–596.

Chao, S., Ning, Y., and Liu, H. (2015). On high dimensional post-regularization prediction intervals. Technical report, arXiv.

Chapelle, O., Schölkopf, B., and Zien, A. (2010). Semi-Supervised Learning. The MIT Press, 1st edition.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm for balanced sampling. Comput. Stat., 21(1):53–62.

Chazal, F., de Vin Silva, and Oudot, S. (2014). Persistence stability for geometric complexes. Geometriae Dedicata, 173(1):193–214.

Chazal, F., Guibas, L. J., Oudot, S. Y., and Skraba, P. (2013). Persistence-based clustering in riemannian manifolds. J. ACM, 60(6).

Chen, H. and Wang, Y. (2011). A penalized spline approach to functional mixed effects model analysis. Biometrics, 67(3):861–870.

Chen, X., Yuan, G., Nie, F., and Huang, J. Z. (2017). Semi-supervised feature selection via rescaled linear regression. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, volume 2017, pages 1525–1531.

Chen, Y., Zhang, L., Wang, H., and E, W. (2020). Deepks: A comprehensive data-driven approach toward chemically accurate density functional theory. J. Chem. Theory Comput., 17(1):170–181.

Cheng, W., Dryden, I. L., and Huang, X. (2016). Bayesian registration of functions and curves. Bayesian Anal., 11(2):447–475.

Chickering, D. M. (2002). Learning equivalence classes of bayesian-network structures. Journal of Machine Learning Research, 2:445–498.

Chipman, H. A., George, E. I., and Mcculloch, R. E. (2010). Bart: Bayesian additive regression trees. Annals of Applied Statistics, pages 266–298.

Chittineni, C. (1980). Learning with imperfectly labeled patterns. Pattern Recognition, 12(5):281–291.

Chu, T. and Glymour, C. (2008). Search for additive nonlinear time series causal models. Journal of Machine Learning Research, 9:967–991.

Claeskens, G., Devijver, E., and Gijbels, I. (2021). Nonlinear mixed effects modeling and warping for functional data using B-splines. Electronic Journal of Statistics, 15(2):5245 – 5282.

Claeskens, G., Silverman, B. W., and Slaets, L. (2010). A multiresolution approach to time warping achieved by a Bayesian prior-posterior transfer fitting strategy. Journal of the Royal Statistical Society Series B, 72(5):673–694.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. Ann. Stat.

Cohen, A. J., Mori-Sánchez, P., and Yang, W. (2012). Challenges for density functional theory. Chem. Rev., 112(1):289–320. PMID: 22191548.

Cohn, D., Ghahramani, Z., and Jordan, M. (1994). Active learning with statistical models. In Advances in Neural Information Processing Systems, volume 7.

Colby, S., McClure, R., Overall, C., Renslow, R., and Mcdermott, J. (2018). Improving network inference algorithms using resampling methods. BMC Bioinformatics, 19.

Colombo, D. and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. Journal of Machine Learning Research, 15(116):3921–3962.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. Annals of Statistics, 40(1):294–321.

Cook, D. (2007). Fisher lecture: Dimension reduction in regression. Statistical Science, 22(1):1–26.

Curtiss, L. A., Raghavachari, K., Redfern, P. C., and Pople, J. A. (1997). Assessment of gaussian-2 and density functional theories for the computation of enthalpies of formation. J. Chem. Phys., 106(3):1063–1079.

da Rosa, J. C., Veiga, A., and Medeiros, M. C. (2008). Tree-structured smooth transition regression models. Computational Statistics and Data Analysis, 58:2469–2488.

Dasgupta, S. and Hsu, D. (2008). Hierarchical sampling for active learning. In Proceedings of the 25th International Conference on Machine Learning, ICML '08, page 208–215, New York, NY, USA. Association for Computing Machinery.

de Mendonça, J. P. A., Mariano, L. A., Devijver, E., Jakse, N., and Poloni, R. (2023). Artificial neural network-based density functional approach for adiabatic energy differences in transition metal complexes. Journal of Chemical Theory and Computation, 19(21):7555–7566. PMID: 37843492.

Deleforge, A., Forbes, F., and Horaud, R. (2015). High-dimensional regression with gaussian mixtures and partially-latent response variables. Statistics and Computing, 25(5):893–911.

Demir, H., Daglar, H., Gulbalkan, H. C., Aksu, G. O., and Keskin, S. (2023). Recent advances in computational modeling of mofs: From molecular simulations to machine learning. Coordination Chemistry Reviews, 484:215112.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, 39(1):1–38.

Derbeko, P., El-Yaniv, R., and Meir, R. (2004). Explicit learning curves for transduction and application to clustering and compression algorithms. Journal of Artificial Intelligence Research, 22(1):117–142.

Devijver, E. (2015a). Finite mixture regression: A sparse variable selection by model selection for clustering. Electronic Journal of Statistics, 9:2642–2674.

Devijver, E. (2015b). An l1-oracle inequality for the lasso in multivariate finite mixture of multivariate gaussian regression models. ESAIM: PS, 19:649–670.

Devijver, E. (2017a). Joint rank and variable selection for parsimonious estimation in a high-dimensional finite mixture regression model. Journal of Multivariate Analysis, 157:1–13.

Devijver, E. (2017b). Model-based regression clustering for high-dimensional data: Application to functional data. Advances in Data Analysis and Classification, 11(2):243–279.

Devijver, E. and Gallopin, M. (2018). Block-diagonal covariance selection for high-dimensional gaussian graphical models. Journal of the American Statistical Association, 113(521):306–314.

Devijver, E., Gallopin, M., and Molinier, R. (2024). Stability of network inference through hierarchical clustering. arXiv.

Devijver, E., Goude, Y., and Poggi, J. (2020). Clustering electricity consumers using high-dimensional regression mixture models. Applied Stochastic Models in Business and Industry, 36(1):159–177.

Devijver, E. and Perthame, E. (2020). Prediction regions through inverse regression. Journal of Machine Learning Research, 21(113):1–24.

Devijver, E. and Samson, A. (2024). Should we correct the bias in confidence bands for repeated functional data? arXiv.

Dick, S. and Fernandez-Serra, M. (2021). Highly accurate and constrained density functional obtained with differentiable programming. Phys. Rev. B, 104:L161109.

Didelez, V. (2003). Graphical models for stochastic processes, pages 138–140.

Ding, M., Flaig, R. W., Jiang, H.-L., and Yaghi, O. M. (2019). Carbon capture and conversion using metal–organic frameworks and mof-based materials. Chem. Soc. Rev., 48:2783–2828.

Dinh, V. C. and Ho, L. S. (2020). Consistent feature selection for analytic deep neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, Advances in Neural Information Processing Systems, volume 33, pages 2420–2431. Curran Associates, Inc.

Diquigiovanni, J., Fontana, M., and Vantini, S. (2022). Conformal prediction bands for multivariate functional data. Journal of Multivariate Analysis, 189:104879.

Domingo, A., Àngels Carvajal, M., and De Graaf, C. (2010). Spin crossover in fe(ii) complexes: An ab initio study of ligand σ-donation. Int. J. Quantum Chem., 110(2):331–337.

Dupuy, J.-F., Loubes, J.-M., and Maza, E. (2011). Non parametric estimation of the structural expectation of a stochastic increasing function. Statistics and Computing, 21(1):121–136.

Eichler, M. and Didelez, V. (2007). Causal reasoning in graphical time series models. In Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, UAI'07, page 109–116, Arlington, Virginia, USA. AUAI Press.

Elith, J., Leathwick, J. R., and Hastie, T. (2008). A working guide to boosted regression trees. Animal Ecology, 77(4):802–813.

Elmi, A., Ratcliffe, S., Parry, S., and Guo, W. (2011). A B-spline based semiparametric nonlinear mixed effects model. Journal of Computational and Graphical Statistics, 20(2):492–509.

Elreedy, D., F. Atiya, A., and I. Shaheen, S. (2019). A novel active learning regression framework for balancing the exploration-exploitation trade-off. Entropy, 21(7).

Faken, D. and Jónsson, H. (1994). Systematic analysis of local atomic structure combined with 3d computer graphics. Computational Materials Science, 2:279–286.

Fan, J. and Li, J. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456):1348–1360.

Fasy, B. T. et al. (2014). Confidence sets for persistence diagrams. Annals of Statistics, 42.

Feofanov, V., Devijver, E., and Amini, M.-R. (2022). Wrapper feature selection with partially labeled data. Applied Intelligence, 52(11):12316–12329.

Feofanov, V., Devijver, E., and Amini, M.-R. (2024). Multi-class probabilistic bounds for majority vote classifiers with partially labeled data. Journal of Machine Learning Research, 25(104):1–47.

Ferraty, F. and Vieu, P. (2006). Nonparametric Functional Data Analysis. Theory and Practice. Springer Series in Statistics. Springer.

Fralick, S. (1967). Learning to recognize patterns without a teacher. IEEE Transactions on Information Theory, 13(1):57–64.

Frazee, A. C., Langmead, B., and Leek, J. T. (2011). ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinformatics, 12(449).

Frei, S., Zou, D., Chen, Z., and Gu, Q. (2022). Self-training converts weak learners to strong learners in mixture models. In International Conference on Artificial Intelligence and Statistics, pages 8003–8021. PMLR.

Frenzel, S. and Pompe, B. (2007). Partial mutual information for coupling analysis of multivariate time series. Physical review letters, 99:204101.

Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). Annals of Statistics, 19(1):1–141.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3):432–441.

Friedman, J. H. (2000). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29:1189–1232.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007). Kernel measures of conditional dependence. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, Advances in Neural Information Processing Systems, volume 20. Curran Associates, Inc.

Fumanal, M., Capano, G., Barthel, S., Smit, B., and Tavernelli, I. (2020). Energy-based descriptors for photo-catalytically active metal–organic framework discovery. J. Mater. Chem. A, 8:4473–4482.

Furmanchuk, A., Agrawal, A., and Choudhary, A. (2016). Predictive analytics for crystalline materials: bulk modulus. RSC Adv, 6(95246).

Furness, J. W., Kaplan, A. D., Ning, J., Perdew, J. P., and Sun, J. (2020). Accurate and numerically efficient r2scan meta-generalized gradient approximation. J. Phys. Chem. Lett., 11(19):8208–8215.

Furness, J. W. and Sun, J. (2019). Enhancing the efficiency of density functionals with an improved iso-orbital indicator. Phys. Rev. B, 99(4):041119.

Galvani, M., Torti, A., Menafoglio, A., and Vantini, S. (2021). Funcc: A new bi-clustering algorithm for functional data with misalignment. Computational Statistics & Data Analysis, 160:107219.

Gamonal, A., Sun, C., Mariano, A. L., Fernandez-Bartolome, E., Guerrero-SanVicente, E., Vlaisavljevich, B., Castells-Gil, J., Marti-Gastaldo, C., Poloni, R., Wannemacher, R., Cabanillas-Gonzalez, J., and Sanchez Costa, J. (2020). Divergent adsorption-dependent luminescence of amino-functionalized lanthanide metal-organic frameworks for highly sensitive NO2 sensors. J. Phys. Chem. Lett., 11(9):3362–3368.

Gao, W., Kannan, S., Oh, S., and Viswanath, P. (2017). Estimating mutual information for discrete-continuous mixtures. Advances in neural information processing systems, 30.

Genovese, C. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. The Annals of Statistics, 28(4):1105–1127.

George, E. I. and Oman, S. (1996). Multiple-shrinkage principal component regression. The Statistician, 45:111–124.

Gervini, D. and Carter, P. A. (2014). Warped functional analysis of variance. Biometrics, 70(3):526–535.

Gervini, D. and Gasser, T. (2004). Self-modelling warping functions. Journal of the Royal Statistical Society: Series B, 66(4):959–971.

Giraud, C. (2008). Estimation of Gaussian graphs by model selection. Electronic Journal of Statistics, 2:542–563.

Goetz, J., Tewari, A., and Zimmerman, P. (2018). Active learning for non-parametric regression using purely random trees. In Advances in Neural Information Processing Systems, volume 31.

Goldberg, D. E. and Holland, J. H. (1988). Genetic algorithms and machine learning. Machine Learning, 3(2):95–99.

Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3):424–38.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007). A kernel statistical test of independence. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, Advances in Neural Information Processing Systems, volume 20. Curran Associates, Inc.

Grün, B. and Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. Computational Statistics & Data Analysis, 51(11):5247–5252.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. Biometrika, 98(1):1–15.

Guo, W. (2002). Functional mixed effects models. Biometrics, 58(1):121–128.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar):1157–1182.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). A Distribution-Free Theory of Nonparametric Regression. Springer.

Hadjadj, L., Devijver, E., Molinier, R., and Amini, M.-R. (2024). Efficient initial data selection and labeling for multi-class classification using topological analysis. In EbookVolume 392: ECAI 2024, Frontiers in Artificial Intelligence and Applications.

Hadjipantelis, P. Z., Aston, J. A. D., Müller, H.-G., and Moriarty, J. (2014). Analysis of spike train data: A multivariate mixed effects model for phase and amplitude. Electron. J. Statist., 8(2):1797–1807.

Hafner, J. (2008). Ab-initio simulations of materials using vasp: Density-functional theory and beyond. J. Comput. Chem, 29:2044–2078.

Hahn, P., Murray, J., and Carvalho, C. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). Bayesian Analysis, 15(3):965–1056. Publisher Copyright: © 2020 International Society for Bayesian Analysis.

Hall, P. (1991). On convergence rates of suprema. Probab Theory Related Fields.

Han, Y., Park, K., and Lee, Y.-K. (2011). Confident wrapper-type semi-supervised feature selection using an ensemble classifier. In 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), pages 4581–4586. IEEE.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). The Elements of Statistical Learning. Springer New York Inc, New York, NY, USA.

Haury, A.-C., Mordelet, F., Vera-Licona, P., and Vert, J.-P. (2012). Tigress: Trustful inference of gene regulation using stability selection. BMC Systems Biology, 6(1):145.

Helland, I. (1992). Maximum likelihood regression on relevant components. Journal of the Royal Statistical Society, Series B, 54:637–347.

Helland, I. and Almøy, T. (1994). Comparison of prediction methods when only a few components are relevant. Journal of the American Statistical Association, 89:583–591.

Hirata, A., Wada, T., Obayashi, I., and Hiraoka, Y. (2020). Structural changes during glass formation extracted by computational homology with machine learning. Commun. Mater, 1:1–4.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67.

Holzmüller, D., Zaverkin, V., KÃ¤stner, J., and Steinwart, I. (2023). A framework and benchmark for deep batch active learning for regression. Journal of Machine Learning Research, 24(164):1–81.

Hong, S. and Kim, D. (2019). Medium-range order in amorphous ices revealed by persistent homology. J. Phys. Condens. Matter, 31.

Hoshikawa, T. (2013). Mixture regression for observational data, with application to functional regression models. Available: http://arxiv.org/abs/1307.0170.

Hu, R., Namee, B. M., and Delany, S. J. (2010). Off to a good start: Using clustering to select the initial training set in active learning. In FLAIRS.

Hume, D. (1738). A Treatise of Human Nature. Oxford University Press.

Hyodo, M., Shutoh, N., Nishiyama, T., and Pavlenko, T. (2015). Testing block-diagonal covariance structure for high-dimensional data. Statistica Neerlandica, 69(4):460–482.

Hyvärinen, A., Shimizu, S., and Hoyer, P. O. (2008). Causal modelling combining instantaneous and lagged effects: An identifiable model based on non-gaussianity. In Proceedings of the 25th International Conference on Machine Learning, ICML '08, pages 424–431, New York, NY, USA. ACM.

Hébrail, G., Hugueney, B., Lechevallier, Y., and Rossi, F. (2010). Exploratory analysis of functional data via clustering and optimal segmentation. Neurocomputing, 73(7):1125–1141. Advances in Computational Intelligence and Learning.

Imbalzano, G., Anelli, A., Giofré, D., Klees, S., Behler, J., and Ceriotti, M. (2018). Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. The Journal of Chemical Physics, 148(24):241730.

Irsoy, O., Yildiz, O. T., and Alpaydin, E. (2012). Soft decision trees. In International Conference on Pattern Recognition.

Jablonka, K. M., Jothiappan, G. M., Wang, S., Smit, B., and Yoo, B. (2021). Bias free multiobjective active learning for materials design and discovery. Nature Communications, 12(1).

Jablonka, K. M., Rosen, A. S., Krishnapriyan, A. S., and Smit, B. (2023). An ecosystem for digital reticular chemistry. ACS Central Science, 9(4):563–581.

Jacobs, R., Jordan, M., Nowlan, S., and Hinton, G. (1991). Adaptive mixtures of local experts. Neural Computation, 3(1):79–87.

Jacques, J. and Preda, C. (2013). Funclust: A curves clustering method using functional random variables density approximation. Neurocomputing, 112:164–171. Advances in artificial neural networks, machine learning, and computational intelligence.

Jacques, J. and Preda, C. (2014). Functional data clustering: A survey. Advances in Data Analysis and Classification, 8:231–255.

Jakse, N., Le Bacq, O., and Pasturel, A. (2004). Prediction of the local structure of liquid and supercooled tantalum. Phys. Rev. B, 70:174203.

Jakse, N., Sandberg, J., Granz, L. F., Saliou, A., Jarry, P., Devijver, E., Voigtmann, T., Horbach, J., and Meyer, A. (2022). Machine learning interatomic potentials for aluminium: application to solidification phenomena. Journal of Physics: Condensed Matter, 35(3):035402.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. Journal of the American Statistical Association, 98(462):397–408.

Janet, J. P. and Kulik, H. J. (2017). Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships. J. Phys. Chem. A, 121(46):8939–8954.

Janková, J. and van de Geer, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. Electron. J. Statist., 9(1):1205–1229.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. J. Mach. Learn. Res., 15(1):2869–2909.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. Ann. Math. Statist., 40(2):633–643.

Jiang, B., Wu, X., Yu, K., and Chen, H. (2019). Joint semi-supervised feature selection and classification through bayesian approach. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):3983–3990.

Jiang, Y., Conglian, Y., and Qinghua, J. (2018). Model selection for the localized mixture of experts models. Journal of Applied Statistics, 45(11):1994–2006.

Jitkrittum, W., Szabó, Z., and Gretton, A. (2017). An adaptive test of independence with analytic kernel embeddings. In Precup, D. and Teh, Y. W., editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 1742–1751. PMLR.

Johnson, E. M., Ilic, S., and Morris, A. J. (2021). Design strategies for enhanced conductivity in metal-organic frameworks. ACS Cent. Sci., 7(3):445–453.

Jones, R. O. (2015). Density functional theory: Its origins, rise to prominence, and future. Rev. Mod. Phys., 87:897–923.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the em algorithm. Neural Comput., 6(2):181–214.

Jose, A., Devijver, E., Jakse, N., and Poloni, R. (2024). Informative training data for efficient property prediction in metal–organic frameworks by active learning. Journal of the American Chemical Society, 146(9):6134–6144. PMID: 38404041.

Jose, A., Mendonça, J., Devijver, E., Jakse, N., Monbet, V., and Poloni, R. (2023). Regression tree-based active learning. Data Mining and Knowledge Discovery, pages 1–41.

Josse, J. and Holmes, S. (2016). Measuring multivariate association and beyond. Statistics Surveys, 10(none):132 – 167.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. Journal of Machine Learning Research, 8:613–636.

Kang, Y., Park, H., Smit, B., and Kim, J. (2023). A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. Nat. Mach. Intell., 5(3):309–318.

Kasim, M. F. and Vinko, S. M. (2021). Learning the exchange-correlation functional from nature with fully differentiable density functional theory. Phys. Rev. Lett., 127:126403.

Kay, S. M. (1993). Fundamentals of statistical signal processing. Prentice Hall PTR.

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In Proceedings of ICNN'95-International Conference on Neural Networks, volume 4, pages 1942–1948. IEEE.

Keribin, C. (2000). Consistent estimation of the order of mixture models. Sankhya: The Indian Journal of Statistics, Series A, 62(1):49–66.

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. Journal of the American Statistical Association, 102(479):1025–1038.

Khalili, A. and Lin, S. (2013). Regularization in finite mixture of regression models with diverging number of parameters. Biometrics, 69(2):436–446.

Kirkpatrick, J., McMorrow, B., Turban, D. H., Gaunt, A. L., Spencer, J. S., Matthews, A. G., Obika, A., Thiry, L., Fortunato, M., Pfau, D., et al. (2021). Pushing the frontiers of density functionals by solving the fractional electron problem. Science, 374(6573):1385–1389.

Kneip, A. and Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. Ann. Statist., 20(3):1266–1305.

Kokoszka, P. and Reimherr, M. (2017). Introduction to Functional Data Analysis. Chapman & Hall / CRC numerical analysis and scientific computing. CRC Press.

Kottke, D., Herde, M., Minh, T. P., Benz, A., Mergard, P., Roghman, A., Sandrock, C., and Sick, B. (2021). scikit-activeml: A Library and Toolbox for Active Learning Algorithms. Preprints.

Kozachenko, L. F. and Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. Problemy Peredachi Informatsii, 23(2):9–16.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. Physical review. E, Statistical, nonlinear, and soft matter physics, 69(6):066138.

Krivobokova, T., Kneib, T., and Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators. Journal of the American Statistical Association, 105(490):852–863.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Systems Biology, 5(1):21.

Lacasse, A., Laviolette, F., Marchand, M., Germain, P., and Usunier, N. (2007). Pac-bayes bounds for the risk of the majority vote and the variance of the gibbs classifier. In Advances in Neural Information Processing Systems, pages 769–776.

Lauritzen, S. L. (1996). Graphical Models. Clarendon Press.

Laviolette, F., Morvant, E., Ralaivola, L., and Roy, J.-F. (2017). Risk upper bounds for general ensemble methods with an application to multiclass classification. Neurocomputing, 219:15–25.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. Signal Processing, 85(4):717 – 736.

Lechner, W. and Dellago, C. (2008). Accurate determination of crystal structures based on averaged local bond order parameters. The Journal of Chemical Physics, 129(11470):7.

Lee, B.-J., Shim, J.-H., and Baskes, M. I. (2003). Semiempirical atomic potentials for the fcc metals cu, ag, au, ni, pd, pt, al, and pb based on first and second nearest-neighbor modified embedded atom method. Phys. Rev. B, 68:144112.

Lee, J., Farha, O. K., Roberts, J., Scheidt, K. A., Nguyen, S. T., and Hupp, J. T. (2009). Metal–organic framework materials as catalysts. Chem. Soc. Rev., 38:1450–1459.

Lee, J., Sun, D., Sun, Y., and Taylor, J. (2016). Exact post-selection inference, with application to the lasso. Ann. Statist., 44(3):907–927.

Lemhadri, I., Ruan, F., Abraham, L., and Tibshirani, R. (2021). Lassonet: A neural network with feature sparsity. Journal of Machine Learning Research, 22(127):1–29.

Levy, M. and Perdew, J. P. (1985). Hellmann-feynman, virial, and scaling requisites for the exact universal density functionals. shape of the correlation potential and diamagnetic susceptibility for atoms. Phys. Rev. A, 32(4):2010.

Li, H., Eddaoudi, M., O'Keeffe, M., and Yaghi, O. M. (1999). Design and synthesis of an exceptionally stable and highly porous metal-organic framework. Nature, 402(6759):276–279.

Li, H., Wang, K., Sun, Y., Lollar, C. T., Li, J., and Zhou, H.-C. (2018). Recent advances in gas storage and separation using metal–organic frameworks. Materials Today, 21(2):108–121.

Li, J.-R., Kuppler, R. J., and Zhou, H.-C. (2009). Selective gas adsorption and separation in metal–organic frameworks. Chem. Soc. Rev., 38:1477–1504.

Li, K. (1991). Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 86(414):316–327.

Li, L., Hoyer, S., Pederson, R., Sun, R., Cubuk, E. D., Riley, P., Burke, K., et al. (2021). Kohn-sham equations as regularizer: Building prior knowledge into machine-learned physics. Physical Review Letters, 126(3):036401.

Li, Y., Qiu, Y., and Xu, Y. (2022). From multivariate to functional data analysis: Fundamentals, recent developments, and emerging areas. Journal of Multivariate Analysis, 188:104806. 50th Anniversary Jubilee Edition.

Liebl, D. and Reimherr, M. (2023). Fast and fair simultaneous confidence bands for functional parameters. Journal of the Royal Statistical Society Series B: Statistical Methodology, 85(3):842–868.

Lim, C. and Yu, B. (2016). Estimation Stability With Cross-Validation (ESCV). Journal of Computational and Graphical Statistics, 25(2):464–492.

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. Journal of the American Statistical Association, 113(522):626–636.

Linero, A. R. and Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(5):1087–1110.

Liu, X. and Yang, M. C. (2009). Simultaneous curve registration and clustering for functional data. Computational Statistics & Data Analysis, 53(4):1361–1376.

Liu, Z., Jiang, X., Luo, H., Fang, W., Liu, J., and Wu, D. (2021). Pool-based unsupervised active learning for regression using iterative representativeness-diversity maximization. Pattern Recognition Letters, 142:11–19.

Lloyd-Jones, L., Nguyen, H., and McLachlan, G. J. (2018). A globally convergent algorithm for lasso-penalized mixture of linear regression models. Computational Statistics & Data Analysis, 119:19–38.

Lorenzen, S. S., Igel, C., and Seldin, Y. (2019). On pac-bayesian bounds for random forests. Machine Learning, 108(8-9):1503–1522.

Malinsky, D. and Spirtes, P. (2018). Causal structure learning from multivariate time series in settings with unmeasured confounding. In Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery, volume 92 of Proceedings of Machine Learning Research, pages 23–47, London, UK. PMLR.

Mariano, A. L., Fernández-Blanco, A., and Poloni, R. (2023). Perspective from a Hubbard U-density corrected scheme towards a spin crossover-mediated change in gas affinity. J. Chem. Phys., 159(15):154108.

Marx, A., Yang, L., and van Leeuwen, M. (2021). Estimating conditional mutual information for discrete-continuous mixtures using multi-dimensional adaptive histograms. In Proceedings of the 2021 SIAM International Conference on Data Mining (SDM), pages 387–395. SIAM.

Masegosa, A., Lorenzen, S., Igel, C., and Seldin, Y. (2020). Second order pac-bayesian bounds for the weighted majority vote. Advances in Neural Information Processing Systems, 33.

Massart, P. (2007). Concentration inequalities and model selection. Lecture Notes in Mathematics. Springer, 33, 2003, Saint-Flour, Cantal.

Maugis, C. and Michel, B. (2011). A non asymptotic penalized criterion for Gaussian mixture model selection. ESAIM. Probability and Statistics., 15:41–68.

Maurer, A. (2004). A note on the pac bayesian theorem.

Mazumder, R. and Hastie, T. (2012). Exact covariance thresholding into connected components for large-scale Graphical Lasso. Journal of Machine Learning Research, 13:781–794.

McAllester, D. (2003). Simplified PAC-Bayesian margin bounds. In Schölkopf, B. and Warmuth, M. K., editors, Learning Theory and Kernel Machines, pages 203–215, Berlin, Heidelberg. Springer Berlin Heidelberg.

McAllester, D. A. (1999). Some pac-bayesian theorems. Machine Learning, 37(3):355–363.

McCallum, A. and Nigam, K. (1998). Employing em and pool-based active learning for text classification. In Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98, page 350–358, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). Umap: Uniform manifold approximation and projection. Journal of Open Source Software, 3(29):861.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, UAI-95, pages 403–410, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Meinshausen, N. (2015). Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77(5):923–945.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. The Annals of Statistics, 34(3):1436–1462.

Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. The Annals of Statistics, 34(3):1436–1462.

Mendelev, M., Kramer, M., Becker, C., and Asta, M. (2008). Analysis of semi-empirical interatomic potentials appropriate for simulation of crystalline and liquid al and cu. Philosophical Magazine, 88(12):1723–1750.

Mesner, O. C. and Shalizi, C. R. (2020). Conditional mutual information estimation for mixed, discrete and continuous data. IEEE Transactions on Information Theory, 67(1):464–484.

Michailidis, G. and d'Alché Buc, F. (2013). Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. Mathematical Biosciences, 246(2):326–334.

Mogensen, S. W. and Hansen, N. R. (2022). Graphical modeling of stochastic processes driven by correlated noise. Bernoulli, 28(4):3023 – 3050.

Morvan, M., Devijver, E., Giacofci, M., and Monbet, V. (2021). Prediction of the NASH through penalized mixture of logistic regression models. The Annals of Applied Statistics, 15(2):952 – 970.

Morvant, E., Koço, S., and Ralaivola, L. (2012). Pac-bayesian generalization bound on confusion matrix for multi-class classification. In Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012.

Motta, F. C. (2018). Topological data analysis: Developments and applications. Adv. Nonlinear Geosci., page 369–391.

Mukherjee, K., Dowling, A. W., and Colón, Y. J. (2022). Sequential design of adsorption simulations in metal–organic frameworks. Mol. Syst. Des. Eng., 7:248–259.

Mukherjee, K., Osaro, E., and Colón, Y. J. (2023). Active learning for efficient navigation of multi-component gas adsorption landscapes in a mof. Digital Discovery, 2:1506–1521.

Nagai, R., Akashi, R., and Sugino, O. (2020). Completing density functional theory by machine learning hidden messages from molecules. Npj Comput. Mater., 6(1).

Nagai, R., Akashi, R., and Sugino, O. (2022). Machine-learning-based exchange correlation functional with physical asymptotic constraints. Phys. Rev. Res., 4(1):013106.

Nandy, A., Duan, C., and Kulik, H. J. (2022). Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery. Curr. Opin. Chem. Eng., 36(100778):100778.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. (2013). Learning with noisy labels. In Advances in Neural Information Processing Systems, pages 1196–1204.

Nauta, M., Bucur, D., and Seifert, C. (2019). Causal discovery with attention-based convolutional neural networks. Machine Learning and Knowledge Extraction, 1(1):312–340.

Nicolás Hernández, J. C. and Jacques, J. (2024). Simultaneous predictive bands for functional time series using minimum entropy sets. Communications in Statistics - Simulation and Computation, 0(0):1–25.

Oliver, G. and Perdew, J. (1979). Spin-density gradient expansion for the kinetic energy. Phys. Rev. A, 20(2):397.

Oman, S. (1991). Random calibration with many measurements: An application of stein estimation. Technometrics, 33:187–195.

O'Neill, J., Jane Delany, S., and MacNamee, B. (2017). Model-free and model-based active learning for regression. In Advances in Computational Intelligence Systems, pages 375–386, Cham. Springer International Publishing.

Orhan, I. B., Le, T. C., Babarao, R., and Thornton, A. W. (2023). Accelerating the prediction of CO2 capture at low partial pressures in metal-organic frameworks using new machine learning descriptors. Commun. Chem., 6(1):214.

Osaro, E., Mukherjee, K., and Colón, Y. J. (2023). Active learning for adsorption simulations: Evaluation, criteria analysis, and recommendations for metal–organic frameworks. Industrial & Engineering Chemistry Research, 62(33):13009–13024.

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., and Aragam, B. (2020). Dynotears: Structure learning from time-series data. In Chiappa, S. and Calandra, R., editors, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, volume 108 of Proceedings of Machine Learning Research, pages 1595–1605. PMLR.

Park, H., Kang, Y., and Kim, J. (2023). Enhancing structure–property relationships in porous materials through transfer learning and cross-material few-shot learning. ACS Applied Materials & Interfaces, 15(48):56375–56385. PMID: 37983088.

Pavlenko, T., Björkström, A., and Tillander, A. (2012). Covariance structure approximation via glasso in high dimensional supervised classification. Journal of Applied Statistics, 39(8):1643–1666.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Pearl, J. (1995). Causal diagrams for empirical research. Biometrika, 82(4):669–688.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY, USA.

Perkovic, E. (2020). Identifying causal effects in maximally oriented partially directed acyclic graphs. In Peters, J. and Sontag, D., editors, Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), volume 124 of Proceedings of Machine Learning Research, pages 530–539. PMLR.

Peters, J., Bauer, S., and Pfister, N. (2022). Causal Models for Dynamical Systems, page 671–690. Association for Computing Machinery, New York, NY, USA, 1 edition.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal Inference by using Invariant Prediction: Identification and Confidence Intervals. Journal of the Royal Statistical Society Series B: Statistical Methodology, 78(5):947–1012.

Peters, J., Janzing, D., and Schölkopf, B. (2013). Causal inference on time series using restricted structural equation models. In Advances in Neural Information Processing Systems 26, pages 154–162.

Phung, Q. M., Feldt, M., Harvey, J. N., and Pierloot, K. (2018). Toward highly accurate spin state energetics in first-row transition metal complexes: A combined caspt2/cc approach. J. Chem. Theory Comput., 14(5):2446–2455.

Prado, S. A., Cabrera-Bosquet, L., Grau, A., Coupel-Ledru, A., Millet, E. J., Welcker, C., and Tardieu, F. (2018). Phenomics allows identification of genomic regions affecting maize stomatal conductance with conditional effects of water deficit and evaporative demand. Plant, Cell & Environment, 41(2):314–326.

Radoń, M. (2019). Benchmarking quantum chemistry methods for spin-state energetics of iron complexes against quantitative experimental data. Phys. Chem. Chem. Phys., 21(9):4854–4870.

Rahimzamani, A., Asnani, H., Viswanath, P., and Kannan, S. (2018). Estimators for multivariate information measures in general probability spaces. Advances in Neural Information Processing Systems, 31.

Ramsay, J. and Silverman, B. (2002). Applied Functional Data Analysis: Methods and Case Studies. Springer-Verlag.

Ramsay, J. and Silverman, B. (2005). Functional Data Analysis. Springer-Verlag, second edition.

Ramsey, J., Spirtes, P., and Zhang, J. (2006). Adjacency-faithfulness and conservative causal inference. In Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, UAI'06, page 401–408, Arlington, Virginia, USA. AUAI Press.

Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.-L., and Celeux, G. (2015). Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. Bioinformatics, 31(9):1420–1427.

Reimann, M. and Kaupp, M. (2023). Spin-state splittings in 3d transition-metal complexes revisited: Toward a reliable theory benchmark. J. Chem. Theory Comput., 19(1):97–108.

Ren, E. and Coudert, F.-X. (2023). Enhancing gas separation selectivity prediction through geometrical and chemical descriptors. Chem. Mater., 35(17):6771–6781.

Ren, J., Qiu, Z., Fan, W., Cheng, H., and Yu, P. S. (2008). Forward semi-supervised feature selection. In Washio, T., Suzuki, E., Ting, K. M., and Inokuchi, A., editors, Advances in Knowledge Discovery and Data Mining, pages 970–976, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. (2021). A survey of deep active learning. ACM Comput. Surv., 54(9).

Riis, C., Antunes, F., Hüttel, F. B., Azevedo, C. L., and Pereira, F. C. (2022). Bayesian active learning with fully bayesian gaussian processes. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, Advances in Neural Information Processing Systems.

Robins, J. M., Scheines, R., Spirtes, P., and Wasserman, L. (2003). Uniform consistency in causal inference. Biometrika, 90(3):491–515.

Robinson, R. W. (1977). Counting unlabeled acyclic digraphs. In Little, C. H. C., editor, Combinatorial Mathematics V, pages 28–43, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rosen, A. S., Fung, V., Huck, P., O'Donnell, C. T., Horton, M. K., Truhlar, D. G., Persson, K. A., Notestein, J. M., and Snurr, R. Q. (2022). High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration. Npj Comput. Mater., 8(1).

Rosen, A. S., Iyer, S. M., Ray, D., Yao, Z., Aspuru-Guzik, A., Gagliardi, L., Notestein, J. M., and Snurr, R. Q. (2021). Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. Matter, 4(5):1578–1597.

Ross, B. C. (2014). Mutual information between discrete and continuous data sets. PLOS ONE, 9(2):1–5.

Ross, J. and Dy, J. (2013). Nonparametric mixture of gaussian processes with constraints. Proc. 30th Int. Conf. Mach. Learn., 28:1346–1354.

Rosset, S. and Tibshirani, R. (2019). From fixed-x to random-x regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. Journal of the American Statistical Association, 0(0):1–14.

Roy, N. and McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. In Brodley, C. E. and Danyluk, A. P., editors, ICML, pages 441–448. Morgan Kaufmann.

Ročková, V. and van der Pas, S. (2020). Posterior concentration for bayesian regression trees and forests. Ann. Statist., 48(4):2108–2131.

Runge, J. (2018a). Causal network reconstruction from time series: From theoretical assumptions to practical estimation. Chaos: An Interdisciplinary Journal of Nonlinear Science, 28(7):075310.

Runge, J. (2018b). Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In International Conference on Artificial Intelligence and Statistics, pages 938–947. PMLR.

Runge, J. (2020). Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In Peters, J. and Sontag, D., editors, Proceedings of Machine Learning Research, volume 124, pages 1388–1397. PMLR.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11).

Rütimann, P. and Bühlmann, P. (2009). High dimensional sparse covariance estimation via directed acyclic graphs. Electronic Journal of Statistics, 3(none):1133 – 1160.

Samé, A., Chamroukhi, F., Govaert, G., and Aknin, P. (2011). Model-based clustering and segmentation of time series with changes in regime. Advances in Data Analysis and Classification, 5(4):301–321.

Samé, A. and Govaert, G. (2012). Online time series segmentation using temporal mixture models and bayesian model selection. In 2012 11th International Conference on Machine Learning and Applications, volume 1, pages 602–605.

Sandberg, J., Voigtmann, T., Devijver, E., and Jakse, N. (2024). Feature selection for high-dimensional neural network potentials with the adaptive group lasso. Machine Learning: Science and Technology, 5(2):025043.

Sandberg, J. E., Devijver, E., Jakse, N., and Voigtmann, T. (2022). Adaptive selection of atomic fingerprints for high-dimensional neural network potentials. In Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

Sasaki, K., Okajima, R., and Yamashita, T. (2018). Liquid structures characterized by a combination of the persistent homology analysis and molecular dynamics simulation. AIP Conf. Proc. 0, 20015.

Schmidt, J., Marques, M. R. G., Botti, S., and Marques, M. A. L. (2019). Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials, 5:1–36.

Schreiber, T. (2000). Measuring information transfer. Physical review letters, 85:461–4.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2):461–464.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. The Annals of Statistics, 43(4):1716–1741.

Scott, C. (2015a). A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In Artificial Intelligence and Statistics, pages 838–846.

Scott, D. W. (2015b). Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons.

Scudder, H. (1965). Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory, 11(3):363–371.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. The Annals of Statistics, 48(3):1514–1538.

Shmueli, G. (2010). To explain or to predict? Statistical Science, 25(3):289–310.

Siedlecki, W. and Sklansky, J. (1993). A note on genetic algorithms for large-scale feature selection. In Handbook of pattern recognition and computer vision, pages 88–107. World Scientific.

Singh, H., Misra, N., Hnizdo, V., Fedorowicz, A., and Demchuk, E. (2003). Nearest neighbor estimates of entropy. American journal of mathematical and management sciences, 23(3-4):301–321.

Singh, S. and Póczos, B. (2016). Finite-sample analysis of fixed-k nearest neighbor density functional estimators. Advances in neural information processing systems, 29.

Snyder, J. C., Rupp, M., Hansen, K., Müller, K.-R., and Burke, K. (2012). Finding density functionals with machine learning. Phys. Rev. Lett., 108(25):253002.

Song, L., Smola, A., Gretton, A., Borgwardt, K. M., and Bedo, J. (2007). Supervised feature selection via dependence estimation. In Proceedings of the 24th international conference on Machine learning, pages 823–830.

Sosso, G. C. and *et al* (2016). Crystal nucleation in liquids: Open questions and future challenges in molecular dynamics simulations. Chemical Reviews, 116(12):7078–7116.

Spirtes, P., Glymour, C., and Scheines, R. (2001). Causation, Prediction, and Search. MIT press, 2nd edition.

Städler, N., Bühlmann, P., and van de Geer, S. (2010). ℓ1-penalization for mixture regression models. TEST, 19(2):209–256.

Stillinger, F. H. and T. A. Weber, T. A. (1982). Hidden structure in liquids. Phys. Rev. A, 25:978.

Strait, J., Kurtek, S., Bartha, E., and MacEachern, S. N. (2017). Landmark-constrained elastic shape analysis of planar curves. Journal of the American Statistical Association, 112(518):521–533.

Strobl, E. V., Zhang, K., and Visweswaran, S. (2019). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. Journal of Causal Inference, 7(1).

Stucky, B. and van de Geer, S. (2018). Asymptotic confidence regions for high-dimensional structured sparsity. IEEE Transactions on Signal Processing, 66(8):2178–2190.

Suarez, A. and Lutsko, F. (2003). Globally fuzzy decision trees for classification and regression. Fuzzy sets and systems, 138:221–254.

Sun, J. and Loader, C. R. (1994). Simultaneous confidence bands for linear regression and smoothing. Ann. Statist., 22(3):1328–1345.

Sun, J., Ruzsinszky, A., and Perdew, J. P. (2015a). Strongly constrained and appropriately normed semilocal density functional. Phys. Rev. Lett., 115(3):036402.

Sun, J., Taylor, D., and Bollt, E. (2015b). Causal network inference by optimal causation entropy. SIAM Journal on Applied Dynamical Systems, 14(1):73–106.

Suppes, P. (1970). A Probabilistic Theory of Causality. Amsterdam: North-Holland Pub. Co.

Swart, M. (2008). Accurate spin-state energies for iron complexes. J. Chem. Theory Comput., 4(12):2057–2066.

Syed, F. H., Tahir, M. A., Rafi, M., and Shahab, M. D. (2021). Feature selection for semi-supervised multi-target regression using genetic algorithm. Applied Intelligence, pages 1–24.

Tan, K. M., Witten, D., and Shojaie, A. (2015). The cluster graphical lasso for improved estimation of gaussian graphical models. Computational Statistics & Data Analysis, 85:23–36.

Tan, Y. V. and Roy, J. (2019). Bayesian additive regression trees and the general bart model. Statistics in Medicine, 38(25):5048–5069.

Telschow, F. J. E., Cheng, D., Pranav, P., and Schwartzman, A. (2023). Estimation of expected Euler characteristic curves of nonstationary smooth random fields. The Annals of Statistics, 51(5):2272 – 2297.

ten Wolde, P. R., Ruiz-Montero, M. J., and Frenkel, D. (1995). Numerical evidence for bcc ordering at the surface of a critical fcc nucleus. Physical Review Letter, 75:2714–2717.

Thiemann, N., Igel, C., Wintenberger, O., and Seldin, Y. (2017). A strongly quasiconvex pac-bayesian bound. In International Conference on Algorithmic Learning Theory, pages 466–492. PMLR.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (Series B), 58:267–288.

Torres-García, W., Zhang, W., Runger, G. C., Johnson, R. H., and Meldrum, D. R. (2009). Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: A non-linear model to predict abundance of undetected proteins. Bioinformatics, 25(15):1905–1914.

Tsamardinos, I. and Borboudakis, G. (2010). Permutation testing improves bayesian network learning. In Balcázar, J. L., Bonchi, F., Gionis, A., and Sebag, M., editors, Machine Learning and Knowledge Discovery in Databases, pages 322–337, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tucker, J. D., Wu, W., and Srivastava, A. (2013). Generative models for functional data using phase and amplitude separation. Comput. Stat. Data Anal., 61(C):50–66.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist., 42(3):1166–1202.

Verma, T. and Pearl, J. (1991). Equivalence and synthesis of causal models. In Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, UAI '90, pages 255–270, New York, NY, USA. Elsevier Science Inc.

Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. Electronic Journal of Statistics, 6:38–90.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242.

Wan, Y. W., Allen, G. I., and Liu, Z. (2015). TCGA2STAT: Simple TCGA data access for integrated statistical analysis in R. Bioinformatics, 32(6):952–954.

Wang, C., Liu, D., and Lin, W. (2013). Metal–organic frameworks as a tunable platform for designing functional molecular materials. Journal of the American Chemical Society, 135(36):13222–13234. PMID: 23944646.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika, 94(3):553–568.

Wang, J., Gu, L., and Yang, L. (2022). Oracle-efficient estimation for functional data error distribution with simultaneous confidence band. Computational Statistics & Data Analysis, 167:107363.

Wang, K. and Gasser, T. (1997). Alignment of curves by dynamic time warping. Ann. Statist., 25:1251–1276.

Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics. Wiley Publishing.

Wilbraham, L., Verma, P., Truhlar, D. G., Gagliardi, L., and Ciofini, I. (2017). Multiconfiguration pair-density functional theory predicts spin-state ordering in iron complexes with the same accuracy as complete active space second-order perturbation theory at a significantly reduced computational cost. J. Phys. Chem. Lett., 8(9):2026–2030.

Willett, R., Nowak, R., and Castro, R. (2005). Faster rates in regression via active learning. In Advances in Neural Information Processing Systems, volume 18.

Wilmer, C. E., Leaf, M., Lee, C. Y., Farha, O. K., Hauser, B. G., Hupp, J. T., and Snurr, R. Q. (2011). Large-scale screening of hypothetical metal-organic frameworks. Nat. Chem., 4(2):83–89.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the Graphical Lasso. Journal of Computational and Graphical Statistics, 20(4):892–900.

Wu, D., Lin, C.-T., and Huang, J. (2019). Active learning for regression using greedy sampling. Information Sciences, 474:90–105.

Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., and Sugiyama, M. (2019). Are anchor points really indispensable in label-noise learning? In Advances in Neural Information Processing Systems, pages 6838–6849.

Xie, L. S., Skorupskii, G., and Dincă, M. (2020). Electrically conductive metal–organic frameworks. Chemical Reviews, 120(16):8536–8580. PMID: 32275412.

Xie, T. and Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys. Rev. Lett., 120:145301.

Xu, L., Jordan, M. I., and Hinton, G. E. (1995). An alternative model for mixtures of experts. In Advances in neural information processing systems, pages 633–640.

Xue, B., Zhang, M., Browne, W. N., and Yao, X. (2015). A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation, 20(4):606–626.

Yan, Y., Rosales, R., Fung, G., and Dy, J. G. (2011). Active learning from crowds. In Proceedings of the 28th International Conference on International Conference on Machine Learning, page 1161–1168.

Yang, M., Chen, Y.-J., and Ji, G.-L. (2010). Semi_fisher score: A semi-supervised method for feature selection. In 2010 International Conference on Machine Learning and Cybernetics, volume 1, pages 527–532. IEEE.

Yin, J. and Li, H. (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. The Annals of Applied Statistics, 5(4):2630–2650.

Younossi, Z., Anstee, Q., Marietti, M., Hardy, T., Henry, L., Eslam, M., George, J., and Bugianesi, E. (2018a). Global burden of nafld and nash: trends, predictions, risk factors and prevention. Nature Reviews Gastroenterology & Hepatology, 15:11–20.

Younossi, Z., Loomba, R., Anstee, Q., Rinella, M., Bugianesi, E., Marchesini, G., Neuschwander-Tetri, B., Serfaty, L., Negro, F., Caldwell, S., Ratziu, V., Corey, K., Friedman, S., Abdelmalek, M., Harrison, S., Sanyal, A., Lavine, J., Mathurin, P., Charlton, M., Goodman, Z., Chalasani, N., Kowdley, K., George, J., and Lindor, K. (2018b). Diagnostic modalities for nonalcoholic fatty liver disease, nonalcoholic steatohepatitis, and associated fibrosis. Hepatology, 68(1):349–360.

Yu, C. and Hansen, J. H. L. (2017). Active learning based constrained clustering for speaker diarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(11):2188–2198.

Yu, Q., Lu, X., and Marron, J. S. (2017). Principal nested spheres for time-warped functional data analysis. Journal of Computational and Graphical Statistics, 26(1):144–151.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. Biometrika, 94(1):19–35.

Yuksel, S. E., Wilson, J. N., and Gader, P. D. (2012). Twenty years of mixture of experts. IEEE Transactions on Neural Networks and Learning Systems, 23(8):1177–1193.

Zan, L., Meynaoui, A., Assaad, C. K., Devijver, E., and Gaussier, E. (2022). A conditional mutual information estimator for mixed data and an associated conditional independence test. Entropy, 24(9).

Zhang, C.-H. and Zhang, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):217–242.

Zhang, H., Nai, J., Yu, L., and Lou, X. W. D. (2017). Metal-organic-framework-based materials as platforms for renewable energy and environmental applications. Joule, 1(1):77–107.

Zhang, H., Ravi, S. S., and Davidson, I. (2020). A graph-based approach for active learning in regression. In SDM.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artificial Intelligence, 172(16):1873 – 1896.

Zhang, J. and Spirtes, P. (2002). Strong faithfulness and uniform consistency in causal inference. In Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence, UAI'03, pages 632–639, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI'11, page 804–813, Arlington, Virginia, USA. AUAI Press.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. (2018). Large-scale kernel methods for independence testing. Statistics and Computing, 28(1):113–130.

Zhang, T. and Yu, B. (2005). Boosting with early stopping: Convergence and consistency. The Annals of Statistics, 33(4):1538 – 1579.

Zhao, J., Lu, K., and He, X. (2008). Locality sensitive semi-supervised feature selection. Neurocomputing, 71(10-12):1842–1849.

Zhao, Y., Song, Z., Li, X., Sun, Q., Cheng, N., Lawes, S., and Sun, X. (2016). Metal organic frameworks for energy storage and conversion. Energy Storage Mater., 2:35–62.

Zheng, X., Zheng, P., and Zhang, R.-Z. (2018). Machine learning material properties from the periodic table using convolutional neural networks. Chem. Sci, 9(8426).

Zhou, H.-C., Long, J. R., and Yaghi, O. M. (2012). Introduction to metal-organic frameworks. Chem. Rev., 112(2):673–674.

Zhou, J. and *et al* (2019). Observing crystal nucleation in four dimensions using atomic electron tomography. Nature, 570:500–503.

Zhou, S., Shen, X., and Wolfe, D. (1998). Local asymptotics for regression splines and confidence regions. Ann. Statist.

Zhu, J., Wang, H., Yao, T., and Tsou, B. K. (2008). Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 1137–1144, Manchester, UK.