PSL UNIVERSITÉ PARIS

HABILITATION À DIRIGER DES RECHERCHES

DE L'UNIVERSITÉ PSL

Présentée à l'Université Paris-Dauphine

Réduction de dimension et estimation adaptative en présence de données fonctionnelles

Présentation des travaux par

Angelina Roche

Le 13/02/2024

Discipline

Mathématiques appliquées

Composition du jury :

Vincent RIVOIRARD Université Paris-Dauphine	*Coordinateur*
Hervé CARDOT Université de Bourgogne	*Rapporteur*
Alexander GOLDENSHLUGER University of Haifa	*Rapporteur*
Alexander MEISTER University of Rostock	*Rapporteur*
Fabienne COMTE Université Paris Cité	*Examinatrice*
Marc HOFFMANN Université Paris-Dauphine	*Examinateur*
Nathalie VIALANEIX INRAe Toulouse	*Examinatrice*

Dauphine | PSL UNIVERSITÉ PARIS

Remerciements

Le processus d'écriture de ce manuscrit d'Habilitation à Diriger des Recherches et celui de la préparation de la soutenance associée n'ont pas été des plus faciles. Le plus difficile a été de surmonter une quantité importante de doutes de différentes natures et je souhaiterais donc remercier, en premier lieu, tous ceux (et celles) qui n'ont pas semblé douter et qui m'ont donné l'énergie d'aller au bout du processus. Mes premiers remerciements vont donc Vincent Rivoirard pour avoir accepté de coordonner cette habilitation et également pour toutes les discussions que nous avons eues depuis mon arrivée à Dauphine. Je souhaiterais remercier également Hervé Cardot, Alexander Meister et Alexander Goldenshluger pour avoir accepté de rapporter ce manuscrit et pour le temps que vous y avez consacré. Je remercie également Fabienne Comte, Marc Hoffmann et Nathalie Vialaneix pour avoir accepté – avec enthousiasme il m'a semblé – de faire partie du jury.

Mes remerciements vont également à l'ensemble des collègues avec qui j'ai eu la chance de collaborer scientifiquement ces dernières années. En premier lieu à Gaëlle Chagny, pour toutes nos discussions autant amicales que scientifiques et également à Giuseppe di Benedetto, Valère Bitseki-Penda, Caroline Bérard, Fabienne Comte, Antoine Channarond, Van Hà Hoang, Olga Mula, Judith Rousseau, Robin Ryder, Camille Sabbah, Giulia Sambataro, Ludovica Saccaro, Mathilde Sautreuil et Nicolas Vergne. Merci à Franck Picard et à Vincent Rivoirard pour avoir accepté de co-encadrer deux thèses avec moi ainsi qu'à Ryad Belakhem et à Nassim Bourarach pour avoir accepté d'être encadré par nous et Anouar Meynaoui pour les discussions très enrichissantes que nous avons eues pendant ton post-doctorat. Je souhaite remercier également Victor Panaretos pour sa visite à Dauphine en février 2020 et pour son invitation à Lausanne en 2022 et les discussions très riches que nous avons eues à ce moment-là. Et enfin mes derniers remerciements scientifiques vont à Elodie Brunel et André Mas pour m'avoir proposé un sujet de thèse sur des thématiques passionnantes et pour avoir guidé mes premiers pas dans la recherche. Les discussions scientifique que j'ai eues avec vous tous et toutes m'ont considérablement fait avancer et je vous remercie pour cela.

Je souhaite remercier également l'ensemble des collègues du Ceremade pour l'ambiance agréable qui y règne et qui fait que l'on est content de venir travailler et notamment à César, Isabelle, Anne-Laure, Gilles et Thomas pour le super travail qu'ils effectuent au quotidien.

Enfin, mes derniers remerciements (mais les plus importants) à ma petite famille : à Vincent,

avec qui je traverse l'aventure du quotidien depuis bientôt 20 ans et à nos petits "reloops" Alexis et Hugo, qui font du quotidien une aventure. Enfin, à ma grande famille, notamment à ma mère, qui n'a pas hésité à faire le voyage depuis Montpellier pour écouter sa fille parler de statistique.

Introduction

Scientific context

The work presented in this manuscript starts at the beginning of my Ph.D. in 2011 in Montpellier. The subject of my thesis was in the field of functional data statistics, which is a branch of statistics that studies data that can be modelled as random functions. The aim was to develop an adaptive estimation procedure for the slope parameter in the functional linear model with scalar output. The specificity of the approach compared to the previous works on the subject was that the estimator was defined by projection onto a random basis, which is the basis of Functional Principal Components. Another specificity was to consider a prediction risk, aligning this work more closely with the field of statistical learning.

After these initial works, in collaboration with Gaëlle Chagny, we considered bandwidth selection for kernel estimators for functional data, adapting to the functional framework the work of Goldenshluger and Lepski (2011) which, at the time, was recent.

At the same time, I became interested in experimental design problems in functional spaces, with the aim of optimizing a "black box" function. This work, which is not detailed in this habilitation thesis, was applied to a computational code in collaboration with Michel Marques who at the time was an engineer at the French Atomic Energy Agency (CEA Cadarache).

In September 2014, I joined the University of Paris-Descartes (now University Paris-Cité) as an ATER (Temporary Teaching and Research Assistant). With Fabienne Comte and Gaëlle Chagny, we worked on the estimation of the hazard rate function in a multiplicative censoring model. This work, in a new framework compared to my previous works, raised new questions, in particular about the support of the estimators.

I then joined the Ceremade (University Paris-Dauphine) as an Assistant Professor in September 2015. During these eight years, while remaining within the framework of one or more of the three themes above, I was able to diversify the approaches and frameworks considered. First, following my accidental participation in a working group on Group Lasso in Montpellier, I became interested in variable selection problems in multivariate functional linear models (i.e., involving multiple variables, at least one of which is infinite-dimensional).

We also formed a small working group in Rouen, consisting of Caroline Bérard, Gaëlle Chagny, Antoine Channarond, and Nicolas Vergne, focusing on adaptive estimation issues for genomic

data processing. This project benefited significantly from the important contribution of Mathilde Sautreuil, who completed an M2 internship and an alternance contract funded by the Normandy region, Van Hà Hoang, a postdoctoral researcher funded by the ANR Smiles project, and Florian Lecocq, a research engineer. Hà, whose postdoctoral research was supervised by Gaëlle and Antoine, worked, among other things, on adaptive estimation in a two-component mixture model and the integration of the PCO method (Lacour et al., 2017) into an EM algorithm (work still in progress). At the same time, Gaëlle and I continued to work together on several themes related to adaptive estimation: particularly on the functional linear regression model with functional output with Anouar Meynaoui, whom we co-supervised during his postdoctoral research (also funded by the ANR Smiles), on the adaptive estimation of conditional quantiles with Camille Sabbah, and on the adaptive estimation of density conditional on a functional variable.

In addition to these activities, I discovered with Judith Rousseau and Giuseppe di Benedetto the challenging world of nonparametric Bayesian statistics. We started to work on a problem that I did not know how to approach from a frequentist perspective at the time: adaptive estimation in a functional single-index model. Building on recent work by Naulet and Rousseau (2017), we were able to propose a prior distribution that achieves a posterior convergence rate that we believe is minimax optimal (the proof of the lower bound is still in progress).

Furthermore, with Olga Mula and Robin Ryder, we created the Stat-Num working group at Ceremade, at the interface between Statistics and Numerical Analysis. The scientific exchanges related to this working group led to the Emergences M&M's project (for Measures and Models) funded by the city of Paris and led by Olga Mula. This allowed us to finance, among other things, a project for the 2021 edition of Cemracs devoted to data assimilation and reduced models in high-dimensional spaces. The results from this project are expected to be submitted for publication soon.

Finally, in 2017, together with Franck Picard and Vincent Rivoirard, we started to work on dimension reduction problems. Ryad Belhakem's Ph.D. thesis, co-supervised by the three of us, focused on minimax convergence rates in functional PCA when the data are discretised and noisy, and on sparse estimation in multivariate functional PCA (Belhakem, 2022). He defended his thesis in 2022 and continues his professional career. Since 2022, we have also had the opportunity to supervise a new Ph.D. student, Nassim Bourarach. This collaboration also took a new turn after the scientific visit of Victor Panaretos to Dauphine in early 2020, during which we managed to define an explicit representation of a Karhunen-Loève-type decomposition for point processes. The theoretical results and their application to real data are very promising and open the door to numerous extensions and generalisations.

More recently, with André Mas, we have been working on defining an adaptive estimation procedure in a functional AR(1) model. Considering a dependence framework for functional data is quite natural. Indeed, many data sets studied in the literature, such as electricity consumption data, are time series sampled at certain points in time (days, years, etc.). These data are usually considered as independent, which can be difficult to verify when considering, for example, the

electricity consumption on the day i and that on the day $i+1$. However, dealing with dependent data poses significant challenges and requires specialized tools. This work echoes an earlier collaboration with Valère Bitseki-Penda on adaptive estimation problems in branching Markov chains. In both cases, one of the main challenges is to establish a sufficiently precise concentration inequality for the empirical processes that need to be controlled.

List of publications and ongoing works

The initials in brackets refer to published and submitted works (in red [...]) and ongoing works (in blue [...]).

Submitted Articles

[CMR]. Adaptive nonparametric estimation in the functional linear model with functional output, with Gaëlle Chagny and Anouar Menaoui.

[BPRR]. Minimax estimation of Functional Principal Components from noisy discretized functional data, with Ryad Belhakem, Franck Picard, and Vincent Rivoirard.

Published Articles

[RLasso]. Variable selection and estimation in multivariate functional linear regression via the Lasso, Electronic Journal of Statistics, 17(2), 3357–3405.

[R22]. New perspectives in smoothing: minimax estimation of the mean and principal components of discretized functional data. The Graduate Journal of Mathematics, Special issue in Probability and Statistics, 7 (2), pp. 95 – 107 (2022).

[CCHR22]. Adaptive nonparametric estimation of a component density in a two-class mixture model, Journal of Statistical Planning and Inference, 216, 51–69 (2022), with Gaëlle Chagny, Antoine Channarond, and Van Hà Hoang.

[BR20]. Local bandwidth selection for kernel density estimation in bifurcating Markov chain model, Journal of Nonparametric Statistics, 32(3), 535–562 (2020), with Valère Bitseki-Penda.

[R18]. Local optimization of black-box functions with high or infinite-dimensional inputs. Application to nuclear safety. Computational Statistics, 33(1), 467–485 (2018).

[CCR17]. Adaptive estimation of the hazard rate with multiplicative censoring. Journal of Statistical Planning and Inference, 184, pp. 27–47 (2017), with Gaëlle Chagny and Fabienne Comte.

[BMR16]. Non-asymptotic Adaptive Prediction in Functional Linear Models. Journal of Multivariate Analysis, 143, pp. 208–232 (2016), with Élodie Brunel and André Mas.

[CR16]. Adaptive estimation in the functional nonparametric regression model,Journal of Multivariate Analysis, 146, pp. 105–118 (2016), with Gaëlle Chagny.

[BR15]. Penalized contrast estimation in functional linear models with circular data,Statistics, 6, pp. 1298–1321, (2015), with Élodie Brunel.

[CR14]. Adaptive and minimax estimation of the cumulative distribution function given a functional covariate. Electronic Journal of Statistics, 8, pp. 2352–2404, (2014), with Gaëlle Chagny.

Proceedings

Adaptive and minimax estimation of the cumulative distribution function given a functional covariate. Chagny, G.; Roche, A., Contributions in infinite-dimensional statistics and related topics, 79–84, Esculapio, Bologna, 2014.

Ongoing Work (in order of appearance in the manuscript)

[BPRR]. Minimax estimation of the principal components in the case of smooth functional data, with Nassim Bourarach, Franck Picard, and Vincent Rivoirard.

[PPRR]. Dimension reduction via functional PCA for Point Processes, with Victor Panaretos, Franck Picard, and Vincent Rivoirard.

[MRRSS]. Forecasting with ABC credible intervals in agent-based dynamical systems, with Olga Mula, Robin Ryder, Giulia Sambataro, and Ludovica Saccaro.

[SR]. Multilevel PCA for functional imaging data, with Devan Sohier.

[DRR]. Posterior contraction rates for Functional Single-Index models, with Giuseppe Di Benedetto and Judith Rousseau.

[MR]. Adaptive estimation in the functional AR(1) model, with André Mas.

[CRS]. Adaptive estimation of the conditional quantile, with Gaëlle Chagny and Camille Sabbah.

[BCCHLR]. Data-driven bandwidth selection method for classification in a semi-parametric model, with Caroline Bérard, Gaëlle Chagny, Antoine Channarond, Van Hà Hoang, and Florian Lecocq.

[CR]. Multiplicative kernel estimator of the density conditionally to a functional data, with Gaëlle Chagny.

Outline

This manuscript is structured around three main themes: dimension reduction, which is the subject of Chapter 1; minimax convergence rates in regression for functional data, which is the topic of Chapter 2; and adaptive estimation for functional and dependent data, which is the subject of Chapter 3. For reasons of thematic coherence, the progression is not chronological, and some works appear in one or more chapters. Chapter 1 can be read independently of Chapters 2 and 3.

Chapter 1, dedicated to dimension reduction, bridges recent work from the theses of Ryad Belhakem ([BPRR]) and Nassim Bourarach ([BPRR]), on minimax estimation of the functions composing the functional PCA basis and associated eigenvalues in the presence of noise and discretization. It also connects with earlier work done during my thesis ([BR15], [BMR16]) and with the postdoctoral work of Anouar Meynaoui ([CMR]), where non-trivial results on the risk of these functions had to be established. Additionally, it relates to recent ongoing work on PCA for point processes ([PPRR]).

Chapter 2 focuses on regression models involving at least one functional variable and, in particular, minimax convergence rates obtained in these models. The considered models include functional linear regression models with scalar output ([BR15], [BMR16], [RLasso]) and functional output [CMR], nonparametric models ([CR14], [CR16], [CR]), and the single-index model ([DRR]).

Finally, Chapter 3 focuses on adaptive estimation, with a particular emphasis on two aspects: model selection in function spaces generated by the PCA basis ([BMR16], [CMR]) and bandwidth selection for kernel estimators in the presence of functional ([CR14], [CR16], [CR]) and dependent data ([BR20]).

Contents

Symbols

$\lfloor \alpha \rfloor$	integer part of a real number α
$\lesssim$	$a_n \lesssim b_n$ means that there exists a constant $c > 0$ such that $a_n \leq c b_n$.
$\asymp$	$a_n \asymp b_n$ means that $a_n \lesssim b_n$ and $b_n \lesssim a_n$.
$(\psi_j)_{j \geq 1}$	basis of eigenfunctions of the covariance operator Γ_X associated to a functional data X.
λ_j	eigenvalue associated to ψ_j
$\lvert A \rvert$	cardinal of a set A
$\lvert\lvert\lvert \cdot \rvert\rvert\rvert$	operator norm, $\lvert\lvert\lvert T \rvert\rvert\rvert = \sup_{f \in \mathcal{H}, \lVert f \rVert_{\mathcal{X}} = 1} \lVert Tf \rVert_{\mathcal{X}}$ for all continuous operator T on $\mathcal{X}$
$\mathcal{C}^0([0,1])$	the set of continuous functions $f : [0,1] \to \mathbb{R}$
$\lVert \cdot \rVert_\infty$	uniform norm ($\lVert f \rVert_\infty = \sup_t \lvert f(t) \rvert$)
$\lVert \cdot \rVert_\alpha$	the α-Hölder norm $\lVert f \rVert_\alpha = \sum_{j=1}^{\lfloor \alpha \rfloor} \lVert f^{(j)} \rVert_\infty + \sup_{s \neq t} \frac{\lvert f^{(\lfloor \alpha \rfloor)}(s) - f^{(\lfloor \alpha \rfloor)}(t) \rvert}{\lvert t - s \rvert^{\alpha - \lfloor \alpha \rfloor}}$
$\mathcal{H}_\alpha(I)$	the set of α-Hölder continuous functions $f : I \to \mathbb{R}$
	(i.e. $f \in \mathcal{H}_\alpha(I)$ iff f is $\lfloor \alpha \rfloor$-times differentiable on I and $\lVert f \rVert_\alpha < +\infty$)
$\mathcal{B}(S)$	Borelian σ-field of a metric space S
Π_S	orthogonal projection onto a subspace S
$(t)_+ = \max\{t; 0\}$	positive part of a real number t

Chapter 1

Dimension reduction: from functional data to point processes

Functional data can be defined as data that can be modeled as realisations of random variables taking values in a function space. Thus, dimension reduction naturally plays a central role in functional data statistics. Dimension reduction is usually performed by projection onto an approximation space, which can be either fixed such as those spanned by Fourier, wavelets or splines basis or random (data-driven). Among these, functional Principal Components Analysis (fPCA), which is the generalisation of classical multivariate Principal Component Analysis to functional spaces, plays a central role.

Let X be a random variable in a Hilbert space $\mathcal{X}$ equipped with a scalar product $\langle \cdot, \cdot \rangle_{\mathcal{X}}$, and $m \in \mathbb{N}^*$. The best m-dimensional space S_m^{PCA} to project the data (in the mean squared sense) is the one that minimizes the quantity

$$\mathbb{E}[\|X - \Pi_S X\|_{\mathcal{X}}^2], \tag{1.1}$$

among the m-dimensional subspaces S of $\mathcal{X}$ (the notation Π_S denotes the orthogonal projection onto S). It is related to the diagonalization of the covariance operator

$$\Gamma : f \in \mathcal{X} \mapsto \mathbb{E}[\langle f, X - \mathbb{E}[X] \rangle_{\mathcal{X}} (X - \mathbb{E}[X])],$$

in the sense that

$$S_m^{PCA} = \text{span}\{\psi_1, \dots, \psi_m\},$$

where $\psi_1, \dots, \psi_m$ are eigenfunctions of Γ associated to the m largest eigenvalues (counted with multiplicity).

For simplicity, we assume that X is centered (i.e. $\mathbb{E}[X] = 0$) and that each eigenvalue of

Γ is of multiplicity 1 (i.e. for all eigenvalue λ_j of Γ, $\dim(\mathrm{Ker}(\Gamma - \lambda_j I)) = 1$)[1]. We denote by $(\psi_j, \lambda_j)_{j\geq 1}$ the eigenfunctions/eigenvalues sequence of Γ (sorted such that $(\lambda_j)_{j\geq 1}$ is a non-increasing sequence).

In the case of second-order periodic stationary processes, which has been studied by Comte and Johannes (2010) and in [BR15] under the name of circular functional data, the eigenfunctions $(\psi_j)_{j\geq 1}$ of Γ are known and coïncides with the Fourier basis. This case aside, when the distribution of the process X is unknown, the eigenfunctions of the operator Γ are also unknown but can be estimated.

Section 1.1 is devoted to the theoretical study of the estimators of the ψ_j's and λ_j's in the case of noisy and discretized functional data in subsection 1.1.1, and in the case where the data are fully observed with perturbation theory tools in subsection 1.1.2. An extension to the case of point processes is presented in Section 1.2. Finally, in Section 1.3, we consider integrating in a statistical analysis the Reduced Order Modeling (ROM) method, developed by numerical analysts to obtain credible intervals for prediction of time-series solutions of a system of differential equations.

1.1 Functional Principal Components Analysis

Estimators of the eigenfunctions/eigenvalues sequence $(\widehat{\psi}_j, \widehat{\lambda}_j)_{j\geq 1}$ are usually obtained by diagonalizing an estimator of the covariance operator Γ. The aim is then to obtain theoretical guarantees that these estimators are convergent and that their convergence rates are optimal in a certain sense.

1.1.1 Minimax rate of estimation for noisy and discretized functional data

First, let us assume that we are observing the functional data on a regular grid and that they are (possibly) corrupted by noise. Then the first step is to construct an estimator of the covariance operator Γ from the observations $\{Z_i(t_h), i = 1, \dots, n; h = 0, \dots, p-1\}$, with $t_h = h/(p-1)$ and

$$Z_i(t_h) = X_i(t_h) + \varepsilon_{i,h},$$

with $\{\varepsilon_{i,h}\}_{i=1,\dots,n;h=0,\dots,p-1} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$ and $X_1, \dots, X_n$ a n-sample of random continuous functions (here $\mathcal{X} = \mathcal{C}_0([0,1])$, and $\langle f, g\rangle_{\mathcal{X}} = \int_0^1 f(t)g(t)dt$).

In this model, the empirical covariance operator

$$\widehat{\Gamma} : f \mapsto \frac{1}{n}\sum_{i=1}^{n} \langle X_i, f\rangle_{\mathcal{X}} X_i$$

is not an estimator of Γ, since the X_i's are not observed. To define an estimator of Γ, we choose

[1] The case where there exists eigenvalues with multiplicity strictly larger than one is discussed in Hsing and Eubank (2015, p. 131)

to first reconstruct the X_i's on an orthonormal system of functions $\{\varphi_1, \dots, \varphi_m\}$ as follows

$$\tilde{X}_i(t) = \sum_{j=1}^{m} \widetilde{\langle X_i, \varphi_j \rangle} \varphi_j(t),$$

where $\widetilde{\langle X_i, \varphi_j \rangle} = \frac{1}{p} \sum_{h=0}^{p-1} Z_i(t_h)\varphi_j(t_h)$ is an estimation of $\langle X_i, \varphi_j \rangle$. Once the X_i's are reconstructed, we get an estimator of the covariance kernel $K(s,t) = \mathbb{E}[X(s)X(t)]$, $s,t \in [0,1]$,

$$\widehat{K}_m(s,t) := \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i(t)\tilde{X}_i(s) \quad \text{and} \quad \widehat{\Gamma}_m : f \mapsto \int_0^1 \widehat{K}_m(s,t) f(t) dt,$$

the associated estimator of Γ. With the estimator $\widehat{\Gamma}_m$, we obtain easily estimators of the eigenfunctions/eigenvalues $(\psi_j, \lambda_j)_{j \geq 1}$ of Γ by taking the eigenvalues/eigenfunctions $(\widehat{\psi}_j^{(m)}, \widehat{\lambda}_j^{(m)})_{j \geq 1}$ of $\widehat{\Gamma}_m$, sorted such that $(\widehat{\lambda}_j^{(m)})_{j \geq 1}$ is a non-increasing sequence .

Remark. The computation of these estimators is quite easy for projection estimators. Indeed, $\widehat{\Gamma}_m$ is uniquely represented by the matrix $G_m = \left(\langle \widehat{\Gamma}_m \varphi_j, \varphi_k \rangle \right)_{1 \leq j,k \leq m}$, the eigenvector $\mathbf{v}_j^{(m)} := \left(v_{1,j}^{(m)}, \dots, v_{m,j}^{(m)} \right)^t$ associated to the j-th largest eigenvalue of G_m gives us directly the j-th eigenfunction of $\widehat{\Gamma}$ by the formula $\widehat{\psi}_j^{(m)}(t) = \sum_{k=1}^{m} v_{k,j}^{(m)} \varphi_k(t)$, $t \in [0,1]$.

We assume the following regularity assumption for the functional data X,

$$\mathbb{E}[(X(t) - X(s))^2] \leq L|t-s|^{2\gamma}, \qquad t,s \in [0,1], \tag{1.2}$$

where $\gamma \in (0,1]$ and $L > 0$. This assumption is equivalent to assuming that the covariance kernel K is a bivariate γ-Hölder-continuous function. It is also linked with the regularity of the ψ_j's and the decreasing rate of the eigenvalues. For instance, assuming that for all j such that $\lambda_j > 0$, ψ_j is γ-Hölder continuous, the Karhunen-Loève theorem implies that

$$\mathbb{E}[(X(t) - X(s))^2] = \sum_{j \geq 1} \lambda_j (\psi_j(t) - \psi_j(s))^2 \leq \sum_{j \geq 1} \lambda_j \|\psi_j\|_\gamma |t-s|^{2\gamma}, \qquad t,s \in [0,1].$$

Hence, (1.2) is verified if $\sum_{j \geq 1} \lambda_j \|\psi_j\|_\gamma < +\infty$. The Brownian motion and Brownian bridge satisfy (1.2) with $\gamma = 1/2$ and $L = 1$, fractional Brownian motion with Hurst exponent H and Hurst index L with $\gamma = H$.

We prove in [BPRR] the following result on the minimax rates for the estimation of the first eigenfunction.

Theorem 1 ([BPRR]). Let, for $\gamma \in (0,1]$, $L > 0$, $\mathcal{R}_\gamma(L)$ the set of all distribution functions on $\mathcal{X}$ verifying (1.2).

Lower bound Assume that $\mathrm{rk}(\Gamma) \geq 2$ then there exists a quantity $c(\sigma) > 0$ such that

$$\inf_{\widehat{\psi}_1} \sup_{P_X \in \mathcal{R}_\gamma(L)} \mathbb{E}[\|\widehat{\psi}_1 - \mathrm{sign}(\langle \widehat{\psi}_1, \psi_1 \rangle)\psi_1\|^2] \geq c(\sigma) \left(p^{-2\gamma} + n^{-1}\right),$$

where the infimum is taken over all estimators calculated from the observations $\{Y_i(t_h), i = 1, \ldots, n; h = 0, \ldots, p-1\}$.

Upper bound Assume that there exists $C_4 > 0$ such that

$$\mathbb{E}\left[\left(\sum_{h=0}^{p-1} X(t_h)v_h\right)^4\right] \leq C_4 \mathbb{E}\left[\left(\sum_{h=0}^{p-1} X(t_h)v_h\right)^2\right]^2, \qquad v = (v_0, \ldots, v_{p-1})^t \in \mathbb{R}^p.$$

Then, there exists two quantities $B(L, K, \gamma) > 0$ and $V(K, \sigma, C_4) > 0$ such that for all $j \geq 1$,

$$\inf_{\widehat{\psi}_j} \sup_{P_X \in \mathcal{R}_\gamma(L)} \mathbb{E}[\|\widehat{\psi}_j - \mathrm{sign}(\langle \widehat{\psi}_j, \psi_j \rangle)\psi_j\|^2] \leq 8\delta_j^{-1} \left(\frac{B(L, K, \alpha)}{p^{2\gamma}} + \frac{\sigma^4}{p^2} + \frac{V(K, \sigma, C_4)}{n}\right),$$

with $\delta_1 = \lambda_1 - \lambda_2$, and for any $j \geq 2$, $\delta_j = \min(\lambda_j - \lambda_{j+1}, \lambda_{j-1} - \lambda_j)$ is the gap between two consecutive eigenvalues (spectral gap)

Key steps of the proof. To prove the lower-bound, we first prove separately that

$$\inf_{\widehat{\psi}_1} \sup_{P_X \in \mathcal{R}_\gamma(L)} \mathbb{E}[\|\widehat{\psi}_1 - \mathrm{sign}(\langle \widehat{\psi}_1, \psi_1 \rangle)\psi_1\|^2] \geq cn^{-1}, \tag{1.3}$$

and

$$\inf_{\widehat{\psi}_1} \sup_{P_X \in \mathcal{R}_\gamma(L)} \mathbb{E}[\|\widehat{\psi}_1 - \mathrm{sign}(\langle \widehat{\psi}_1, \psi_1 \rangle)\psi_1\|^2] \geq c' p^{-2\gamma}. \tag{1.4}$$

Inequality (1.3) comes from the properties of the Fourier basis that allows us to construct two functions $\psi_{1,A}$ and $\psi_{1,B}$ such that $\|\psi_{1,A}\| = \|\psi_{1,B}\| = 1$ and $\|\psi_{1,A} - \psi_{1,B}\| \asymp n^{-1}$ and generate from it two samples $\{Y_i^A(t_h), i = 1, \ldots, n; h = 0, \ldots, p-1\} \sim \mathbb{P}_A^{\otimes n}$ and $\{Y_i^B(t_h), i = 1, \ldots, n; h = 0, \ldots, p-1\} \sim P_B^{\otimes n}$ such that the Kullback-Leibler divergence of $\mathbb{P}_B^{\otimes n}$ with $\mathbb{P}_A^{\otimes n}$ is uniformly bounded. For (1.4), we use Assouad's Lemma and follows the general scheme described in Tsybakov (2009). The models we define consists in adding some perturbations of the function $\psi_{1,0} = \mathbf{1}_{[0,1]}$ around each grid point t_h. This allows us to define functions $\psi_{1,\omega}$ such that $\psi_{1,\omega}(t_h) = \psi_{1,0}(t_h)$ for all h and $\|\psi_{1,\omega} - \psi_{1,\omega'}\| \geq p^{-\gamma}$ for all $\omega \neq \omega'$. Then the functions are at distance $p^{-\gamma}$ but generate the same distribution of the observations.

The upper-bound is attained by an histogram estimator with $m = p$ bins i.e.

$$\varphi_m = p^{-1/2}\mathbf{1}_{[t_h; t_{h+1})}, h = 0, \ldots, p-1.$$

The proof of the upper-bound relies on inequalities proven in Bosq (2000) that give an upper-

bound on the distance between the eigenfunctions of two operators by controlling the distance, in operator norm, of these two operators. Remark that the key argument for which the lower and upper-bounds match is that $\gamma \leq 1$, which implies that $\frac{\sigma^2}{p^2} \lesssim \frac{B(L,k,\alpha)}{p^{2\gamma}}$. □

We also obtained a result in probability and an upper-bound on the estimation risk of the eigenvalues.

The fact that the minimax rate is obtained without any regularization or smoothing, even when the noise variance σ^2 is non null, is quite surprising, especially in the field of functional data, where smoothing is a common practice, but also from a non-parametric statistics point of view, since we are used to regularizing estimators. The obtention of similar results in the case $\gamma > 1$ – that is to say, to processes X that are $\lfloor\gamma\rfloor$-times differentiable such that

$$\mathbb{E}\left[\left(X^{(\lfloor\gamma\rfloor)}(t) - X^{(\lfloor\gamma\rfloor)}(s)\right)^2\right] \leq L|t-s|^{\gamma-\lfloor\gamma\rfloor}$$

– is an open question and is now under investigation by Nassim Bourarach who started his PhD in October 2022 under the co-supervision of Franck Picard, Vincent Rivoirard and myself. Nassim has also obtained results for the lower bound on the estimation of the eigenvalues and inconsistency results for the estimation of eigenfunctions in the case where the spectral radius of Γ can be taken arbitrarily small.

Similar rates are obtained for the estimation of the mean of functional data by Cai and Yuan (2011).

It is noteworthy that the fact that the grid $\{t_0, \dots, t_{p-1}\}$ is fixed and not random is crucial. In the case where the grid is random, the convergence rates are completely different and regularization is required to achieve optimal rates (see [R22] and references therein).

1.1.2 Perturbation theory [BMR16] [CMR]

In the example of functional principal components regression that we studied in [BMR16], the aim is to estimate the slope function β^* from observations $\{(X_i, Y_i), i = 1, \dots, n\}$ which are i.i.d. copies of a couple of random variables $(X, Y) \in \mathcal{X} \times \mathbb{R}$, such that

$$Y = \langle \beta^*, X\rangle_{\mathcal{X}} + \varepsilon,$$

with ε a centered noise, independent of X. In this subsection, we consider the ideal case where we observe $X_i(t)$ for all $t \in [0,1]$. In that case, the empirical covariance operator $\widehat{\Gamma}$ is calculable from the observations and our estimators $(\widehat{\psi}_j, \widehat{\lambda}_j)_{j\geq 1}$ are the eigenfunctions/eigenvalues of $\widehat{\Gamma}$.

Let

$$\widehat{S}_m^{PCA} = \text{span}\{\widehat{\psi}_1, \dots, \widehat{\psi}_m\}$$

the space spanned by the m-first elements of the fPCA basis that we use as an approximation

space. The risk, in prediction norm $\|\cdot\|_\Gamma = \|\Gamma^{1/2}\cdot\|$, of the least squares estimator

$$\widehat{\beta}_m \in \arg\min_{\beta\in\widehat{S}_m^{PCA}} \frac{1}{n}\sum_{i=1}^n \left(Y_i - \langle\beta, X_i\rangle_{\mathcal{X}}\right)^2,$$

can be decomposed as follows

$$\begin{aligned}\|\widehat{\beta}_m - \beta^*\|_\Gamma^2 &= \|\Pi_{\widehat{S}_m^{PCA}}\beta^* - \beta^*\|_\Gamma^2 + \|\widehat{\beta}_m - \Pi_{\widehat{S}_m^{PCA}}\beta^*\|_\Gamma^2 \\ &\leq \left(\|\Pi_{\widehat{S}_m^{PCA}}\beta^* - \Pi_{S_m^{PCA}}\beta^*\|_\Gamma + \|\Pi_{S_m^{PCA}}\beta^* - \beta^*\|_\Gamma\right)^2 + \|\widehat{\beta}_m - \Pi_{\widehat{S}_m^{PCA}}\beta^*\|_\Gamma^2.\end{aligned}$$

The term $\|\Pi_{S_m^{PCA}}\beta^* - \beta^*\|_\Gamma^2$ can be seen as a bias term and, if

$$\beta^* \in \mathcal{E}_b(R) = \left\{\beta\in\mathcal{X}, \sum_{j\geq 1} j^{2b}\langle\beta,\psi_j\rangle^2 \leq R^2\right\}$$

then we have easily

$$\|\Pi_{S_m^{PCA}}\beta^* - \beta^*\|_\Gamma = \sum_{j>m}\lambda_j\langle\beta^*,\psi_j\rangle^2 \leq m^{-2b}\sum_{j>m}\lambda_j j^{2b}\langle\beta^*,\psi_j\rangle^2 \lesssim \lambda_m m^{-2b}.$$

The last term $\|\widehat{\beta}_m - \Pi_{\widehat{S}_m^{PCA}}\beta^*\|_\Gamma^2$ can be seen as a variance term and is (under mild assumptions) of order m/n. Then the residual term $\|\Pi_{\widehat{S}_m^{PCA}}\beta^* - \Pi_{S_m^{PCA}}\beta^*\|_\Gamma$ that takes into account the randomness of the fPCA basis should be negligible with respect to the optimal risk $\min_{m\geq 1}\{\lambda_m m^{-2b} + m/n\}$. We use tools from perturbation theory that allows us to prove the following result.

Lemma 2 (Lemma 15 of [BMR16]). If $\lambda_j \asymp j^{-2\gamma}$ for $\gamma > 0$ and $\beta^* \in \mathcal{E}_b(R)$, under some subgaussianity assumption of the scores $\xi_j = \langle X,\psi_j\rangle/\sqrt{\lambda_j}$,

$$\mathbb{E}[\|\Pi_{S_m^{PCA}}\beta^* - \Pi_{\widehat{S}_m^{PCA}}\beta^*\|_\Gamma^2] \leq C_1\frac{\ln^3(m)}{n}m^{\max\{(1-2b)_+,2(1-\gamma-b)\}} + C_2\frac{\ln^9(n)}{n^2}m^{(4+2(\gamma-b)_+-4\gamma)_++2}.$$

Idea of proof. The idea is to write the projector in the form of an integral on a complex path $\boldsymbol{\gamma}$. We consider for instance the contour $\boldsymbol{\gamma} = \cup_{j=1}^m \mathcal{C}_j$ where $\mathcal{C}_j$ is the circle of the complex plane centered at the eigenvalue λ_j of radius $\delta_j/2$ (recall that $\delta_j = \min\{\lambda_j - \lambda_{j+1}; \lambda_{j-1} - \lambda_j\}$ is the difference between two consecutive eigenvalues) represented in Fig. 1.1.

First recall that the index of a complex number z with respect to a closed path $\boldsymbol{\gamma}$ (with $z \neq \boldsymbol{\gamma}(t)$, for all t) writes

$$\mathrm{Ind}_{\boldsymbol{\gamma}}(z) = \frac{1}{2i\pi}\int_{\boldsymbol{\gamma}}\frac{d\zeta}{\zeta - z},$$

and that, for a simple contour like $\boldsymbol{\gamma}$, $\mathrm{Ind}_{\boldsymbol{\gamma}}(z) = 1$ if z is inside one of the circle $\mathcal{C}_j$ for $j \leq m$ and 0 otherwise (see Theorem 10.11 p. 204 of Rudin 1966). By definition of the projector Π_m,

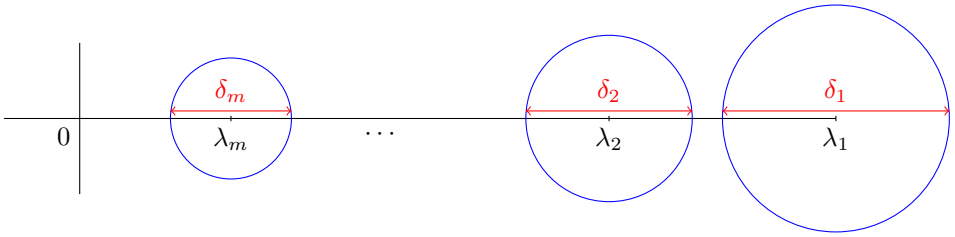

Figure 1.1: Contour made of disjoint circles

for all $j \geq 1$, and using the definition of integrals of continuous functions

$$\Pi_{S_m^{PCA}}\psi_j = \mathbf{1}_{\{j\leq m\}}\psi_j = \frac{1}{2i\pi}\int_{\boldsymbol{\gamma}} \frac{dz}{z-\lambda_j}\psi_j = \frac{1}{2i\pi}\int_{\boldsymbol{\gamma}} \frac{1}{z-\lambda_j}\psi_j dz = \frac{1}{2i\pi}\int_{\boldsymbol{\gamma}} (zI-\Gamma)^{-1}\psi_j dz, \qquad j \geq 1,$$

meaning that[2],

$$\Pi_{S_m^{PCA}} = \frac{1}{2i\pi}\int_{\boldsymbol{\gamma}} (zI-\Gamma)^{-1}dz.$$

Moreover, it can be proved that, under an assumption of sub-gaussianity of the coefficients $\langle X, \psi_j\rangle$, with large probability, $|\widehat{\lambda}_j - \lambda_j| < \delta_j$, for all $j = 1, \ldots, m$ (if m is not too large, but for instance $m \leq n$ works, see Lemmas 11 p. 223 and 13 p. 224 of [BMR16]). Hence we also have

$$\Pi_{\widehat{S}_m^{PCA}} = \frac{1}{2i\pi}\int_{\boldsymbol{\gamma}} (zI-\widehat{\Gamma})^{-1}dz.$$

This leads to the following equality (see Theorem 5.1.4 of Hsing and Eubank 2015 for a precise statement), true with large probability

$$\begin{aligned}\Pi_{S_m^{PCA}} - \Pi_{\widehat{S}_m^{PCA}} = \sum_{j=1}^{m}\sum_{k\neq j}\frac{1}{\lambda_k-\lambda_j}\Big(\pi_j(\widehat{\Gamma}-\Gamma)\pi_k + \pi_k(\widehat{\Gamma}-\Gamma)\pi_j\Big) \\ + \frac{1}{2i\pi}\int_{\boldsymbol{\gamma}} R(z)\sum_{k=2}^{\infty}(-(\widehat{\Gamma}-\Gamma)R(z))^k dz,\end{aligned} \tag{1.5}$$

with $\pi_j(\cdot) = \langle\cdot, \psi_j\rangle\psi_j$ the orthogonal projector onto $\text{span}(\psi_j)$ and $R(z) = (\Gamma - zI)^{-1}$ is the resolvent operator of Γ. This formula explicitly expresses the difference between the two operators Π_{S_m} and $\Pi_{\widehat{S}_m}$ as a function of $\widehat{\Gamma}-\Gamma$, which can be controlled with classical tools such as TCL and Bernstein type inequalities, since $\widehat{\Gamma}$ is a moment estimator. The first term gives the exact order of the difference and the last term can be proven (sometimes under some additional assumptions) to be negligible.

□

Hence this residual term due to the randomness of the basis is negligible compared to the

[2]The integral below is well defined (see Rudin 1966, Definition 3.26).

variance term of order m/n since $\gamma + b > 1/2$. Remark that the upper-bound does not depend on the spectral gap δ_m which is crucial here since it can be very large (for instance, if $\lambda_j = j^{-2\gamma}$, $\delta_j^{-1} \asymp j^{2\gamma+1}$). Finally, this allows us to prove that the estimator $\widehat{\beta}_{m^*}$ achieves the minimax rates described in Table 2.1, p. 23 with appropriate choices of m^*.

1.2 Principal Components Analysis for Point Processes [PPRR]

Another open question is on which objects exactly PCA may be extended. The natural framework to consider PCA is Hilbert spaces, as the Hilbert structure links the optimization problem (1.1) to the diagonalisation of the covariance operator and leads naturally to the series expansion

$$X = \sum_{j \geq 1} \sqrt{\lambda}_j \xi_j \psi_j, \tag{1.6}$$

with the convergence lying in an $\mathbb{L}^2$-sense. Now, for functional data, as a second step, with an additional assumption that the paths of X are mean-square continuous, the Karhunen-Loève theorem (see e.g. Theorem 7.3.5, p. 188 in Hsing and Eubank 2015) states that the convergence (1.6) holds uniformly that is to say the sequence $\Pi_{S_m^{PCA}} X = \sum_{j=1}^m \sqrt{\lambda}_j \xi_j \psi_j$ also converges, when $m \to \infty$, in the Banach space $\mathcal{C}^0([0,1])$ equipped with the norm $\|\cdot\|_\infty$.

We study the problem of extending these results to point processes that are seen as random measures. Carrizo Vergara (2022) proves series expansions of the form (1.6) for signed random measures in $\mathbb{R}^d$ by using the fact that the space of finite random signed measure in $\mathbb{R}^d$ can be injected in the Hilbert space $\mathbb{L}^2(\mathbb{R}^d)$ via the transformation $\mu \mapsto F_\mu$ where $F_\mu(t_1, \ldots, t_d) = \mu([0, t_1] \times \ldots \times [0, t_d])$. The questions that we intend to answer are the following. First we want to obtain a constructive version of the series expansion of Carrizo Vergara (2022) that allows us to define estimators. Second, we aim at obtaining explicit decomposition for some well-known distributions of point processes (Poisson process and Hawkes processes) and third we aim at obtaining equivalent of Mercer and Karhunen-Loève theorem for the associated random measures (with uniform convergence when it is possible).

Assume we observe n i.i.d. point processes $N_1, \ldots, N_n \sim_{i.i.d.} N$. We associate to a process $N = \{t_1, \ldots, t_{|N|}\}$ a random measure on $([0,1], \mathcal{B}([0,1]))$,

$$\Pi(A) = \sum_{j=1}^{|N|} \mathbf{1}_{\{t_j \in A\}} = |A \cap N|, \qquad A \in \mathcal{B}([0,1]), i = 1, \ldots, n.$$

It is natural to define the intensity measure

$$\Lambda(A) = \mathbb{E}[\Pi(A)], \qquad A \in \mathcal{B}([0,1]),$$

which can be seen as a moment of order 1 and we aim at characterizing, as we do naturally for functional data, the variation of each measure Π_i around its mean Λ or, written differently, the

variations of

$$\Delta_i = \Pi_i - \Lambda.$$

$\Delta_1, \ldots, \Delta_n$ is, by definition, a sample a random signed measures. We define the associated covariance measure,

$$C_\Delta(A \times B) = \mathrm{Cov}(\Delta(A), \Delta(B)), \qquad A, B \in \mathcal{B}([0,1]),$$

which defines a measure on the product σ-field $\mathcal{B}([0,1]) \otimes \mathcal{B}([0,1])$ and can be seen as a moment of order 2 of Π.

We define a kernel associated to the measure C_Δ as follows

$$K_\Delta(s,t) = C_\Delta([0,s] \times [0,t]), \qquad s,t \in [0,1],$$

and diagonalize the integral operator associated to C_Δ,

$$\Gamma_\Delta : f \mapsto \int_0^1 K_\Delta(\cdot, t) f(t) dt.$$

The associated eigenfunctions/eigenvalues sequence is denoted by $(\eta_j, \lambda_j)_{j \geq 1}$. It can be proved that the function η_j is a function with bounded variations for all j such that $\lambda_j > 0$. Hence, its derivative, in the distribution sense, is a measure that we denote by μ_j. We obtain the following results.

Theorem 3 ([PPRR]). Assume $\mathbb{E}[\Pi^2([0,1])] < +\infty$ and that K_Δ is a continuous bivariate function.

1. For all function $\varphi : [0,1]^2 \to \mathbb{R}$ of class $\mathcal{C}^2$ and with support in $(0,1)^2$,

$$\sum_{j=1}^{m} \lambda_j \langle \mu_j \otimes \mu_j, \varphi \rangle \xrightarrow[m \to +\infty]{} \langle C_\Delta, \varphi \rangle,$$

   where, for a measure μ and a function φ, we note $\langle \mu, \varphi \rangle = \int \varphi d\mu$ and $\mu_j \otimes \mu_j$ is the product measure of μ_j and μ_j.

2. Moreover, if

$$\sup_{m \geq 1} \left\| \sum_{j=1}^{m} \lambda_j \mu_j \otimes \mu_j \right\|_{TV} < +\infty,$$

   then we have the uniformity result

$$\lim_{m \to \infty} \sup_{\varphi \in \mathcal{C}_0([0,1]^2), \|\varphi\|_\alpha \leq L} \left| \left\langle C_\Delta - \sum_{j=1}^{m} \lambda_j \mu_j \otimes \mu_j, \varphi \right\rangle \right| = 0.$$

3. If, in addition, $t \mapsto \Lambda([0,t])$ is a continuous function on $[0,1]$,

$$\lim_{m\to\infty} \sup_{\varphi\in\mathcal{C}_1(0,1),\|\varphi'\|\leq 1,\varphi(1)=0} \mathbb{E}\left[\left(\langle\Pi_1,\varphi\rangle - \sum_{j=1}^{m}\sqrt{\lambda_j}\xi_j\langle\mu_j,\varphi\rangle\right)^2\right] = 0.$$

Idea of proof. The key of the proof is to apply Mercer's theorem on the kernel K_Δ to obtain 1. and Karhunen-Loève theorem to obtain 3. The uniform result 2. relies on the properties of the total variation distance of a measure as the dual norm on the space of continuous functions $\mathcal{C}_0$ equipped with the uniform norm $\|\cdot\|_\infty$

$$\|\mu\|_{TV} = \sup_{\varphi\in\mathcal{C}_0,\|\varphi\|_\infty\leq 1}\langle\mu,\varphi\rangle.$$

We also consider a regularizing sequence to deduce the uniform convergence result 2. from the weak convergence result 1. □

Theorem 3 can be considered as an extension of the Karhunen-Loève Theorem and Mercer's Theorem to point processes.

Now the following question that we intend to answer is if it is possible to obtain explicit Karhunen-Loève decomposition for some examples of point processes. We considered the following examples.

Homogeneous Poisson process The kernel associated to the homogeneous Poisson process of intensity w,

$$K_\Delta(s,t) = w\min\{s,t\}$$

coincides with the one of the Brownian motion with variance w. Hence the Karhunen-Loève decomposition can be derived directly from the one of the Brownian motion and we get

$$\Pi([0,t]) = wt + \sum_{j\geq 1}\sqrt{\lambda_j}\xi_j\eta_j(t), \qquad t\in[0,1],$$

where

$$\lambda_j = \frac{w}{(j\pi-\pi/2)^2} \text{ and } \eta_j(t) = \sqrt{2}\sin(\pi(2j-1)t/2),$$

and the scores ξ_j can be written

$$\xi_j = \sqrt{2}\sum_{X\in N}\cos\left(\frac{\pi(2j-1)X}{2}\right) + (-1)^j\sqrt{\lambda_j}.$$

In other words, for a borelian set A, $\Pi(A)$ is, up to the multiplicative constant w, equal to the Lebesgue measure of A plus a random perturbation that can be written explicitly. Note that, unlike the Brownian motion, the distribution of the scores is not Gaussian and

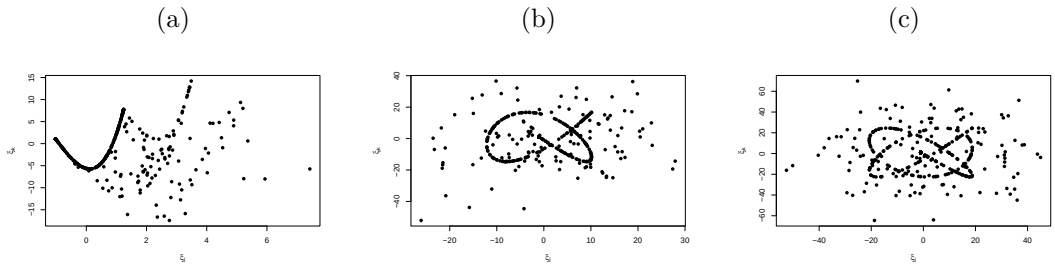

Figure 1.2: Sample of scores (ξ_j, ξ_k) (axis j versus axis k) of homogeneous Poisson processes. (a) $j = 1$, $k = 2$; (b) $j = 3$, $k = 4$; (c) $j = 5$, $k = 6$.

has a particular structure as represented in Figure 1.2.

Inhomogeneous Poisson processes This case is more intricate and, unless the particular case of the homogeneous process considered previously, the eigenfunctions $(\eta_j)_{j\geq 1}$ and eigenvalues $(\lambda_j)_{j\geq 1}$ appearing in the Karhunen-Loève decomposition are not explicit. However, we have established that they are solutions of some second-order differential equations which allows us to derive some properties.

Proposition 4. Let Π be an inhomogeneous Poisson process whose intensity $t \mapsto w(t)$ is positive on $(0, 1)$ and verifies $\int_0^1 w(t)dt < +\infty$.

1. $(\lambda_j, \eta_j)_{j\geq 1}$ are the eigenelements of the operator Γ_Δ iff $(\lambda_j, F_j)_{j\geq 1}$ with $F_j(t) = \int_t^1 \eta_j(t)dt$ are solutions of

$$\begin{cases} -\lambda y''(t) = w(t)y(t), & t \in (0, 1) \\ y(1) = 0, y'(0) = 0. \end{cases} \tag{1.7}$$

2. The eigenvalues λ_j are all distinct and verifies

$$\lambda_j \sim_{j\to+\infty} \frac{\left(\int_0^1 w^{1/2}(t)dt\right)^2}{\pi^2 j^2}.$$

3. The function F_j has exactly $j - 1$ zeros on $(0, 1)$.

Idea of proof. 1. comes by differentiation of the equation $\lambda_j\eta_j = \Gamma_\Delta\eta_j$. 2. and 3. are obtained by identifying Eq. (1.7) with a Sturm-Liouville problem (Zettl, 2005). □

Remark. The case of the homogeneous process corresponds to the case where w is a constant function on $[0, 1]$. In that case,the differential equation (1.7) can be solved explicitly and we obtain the expected solution $(\lambda_j, \eta_j)_{j\geq 1}$ written above.

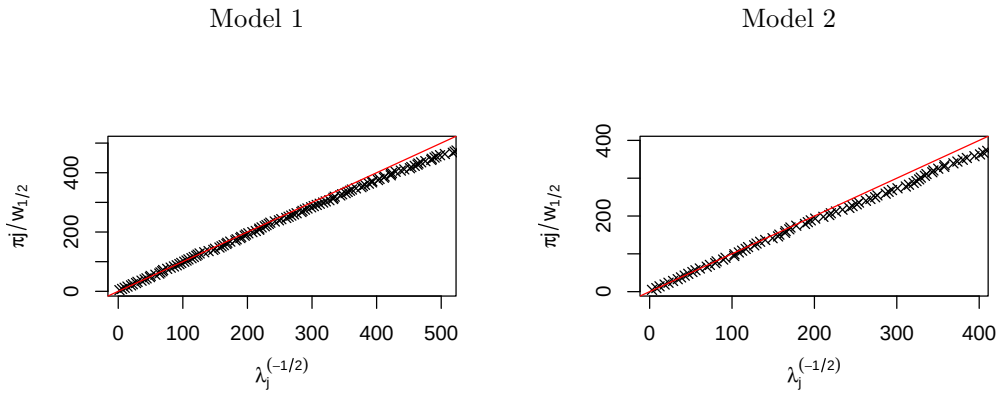

Figure 1.3: Plot of $\left(\pi j/\left(\int_0^1 \sqrt{w(u)}du\right)^2, \widehat{\lambda}_j^{-1/2}\right)$, $j = 1, \ldots, \lfloor \mathrm{rk}\widehat{\Gamma}/2 \rfloor$

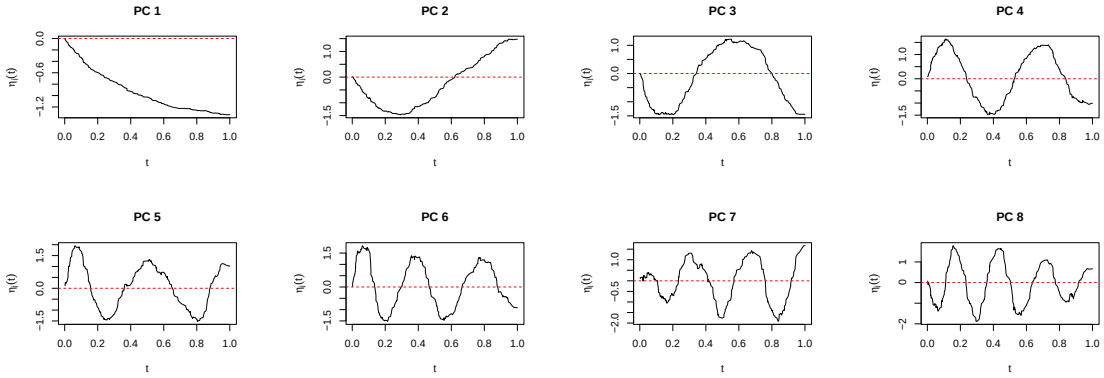

Figure 1.4: Plot of $\widehat{\eta}_j$, $j = 1, \ldots, 8$ for Model 1

On simulated data, we observe the behavior predicted by the theory for two different choices of the function w (model 1: $w(t) = e^{-t}$ and model 2 : $w(t) = t$). Fig. 1.3 and Fig. 1.4 illustrate the behavior of the eigenfunctions/eigenvalues sequences. Note that, unintentionally, we also obtain a method for approximating the solutions of the differential equation (1.7) from the simulation of Poisson processes and the approximation of the associated eigenfunctions/eigenvalues.

Hawkes processes Hawkes processes are particular classes of point processes that allows the possibility of self-excitation of the process. They are used to model for instance neuronal spikes or earthquakes. The intensity function of a univariate Hawkes process can be written

$$w(t) = w_0 + \sum_{X \in N, X \leq t} h(t - X),$$

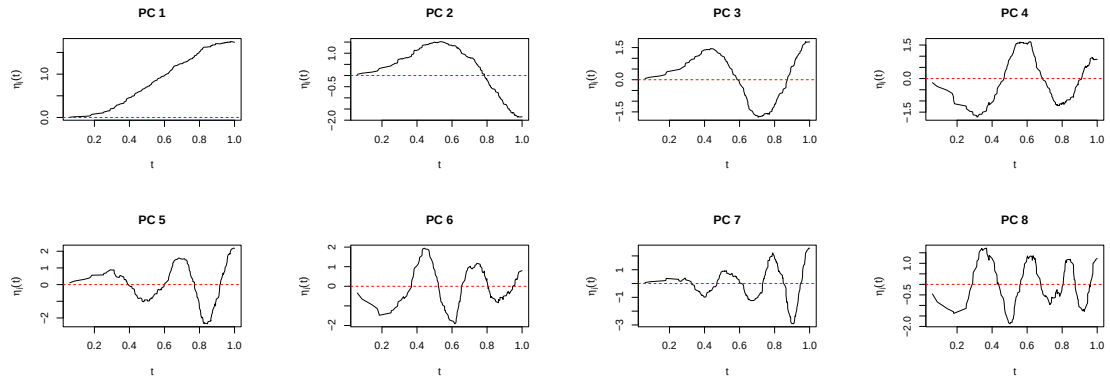

Figure 1.5: Plot of $\widehat{\eta}_j$, $j = 1, \ldots, 8$ for Model 2

with $w_0 > 0$ the baseline intensity and $h : \mathbb{R}_+ \to \mathbb{R}_+$ an exciting function. We obtain some results in the particular case of an intensity function of the form $h(t) = h_0 e^{-\tilde{h}t}$. The existence of a stationary process is ensured under the condition $\int_0^{+\infty} h(t)dt < 1$ meaning that $0 < h_0 < \tilde{h}$ (see e.g. Daley and Vere-Jones 2008, Example 12.5(c)). In the case of Hawkes processes, the eigenvalues/eigenfunctions sequence are obtained by solving a perturbed version of the differential equation associated to the homogeneous Poisson process. Under a condition on the parameters of the process, the following proposition allows us to establish that the eigenvalues and eigenfunctions have a behavior very similar to that of the homogeneous Poisson process.

Proposition 5. Let Π be a stationary Hawkes process whose intensity is $h(t) = h_0 e^{-\tilde{h}t}$ with $0 < h_0 < \tilde{h}$.

1. Then, $(\lambda_j, \eta_j)_{j\geq 1}$ are the eigenelements of the operator Γ_Δ iff $(\lambda_j, F_j)_{j\geq 1}$ with $F_j(t) = \int_t^1 \eta_j(t)dt$ are solutions of

$$\begin{cases} -\lambda y''(t) = wy(t) + \frac{h_0 w(2\tilde{h}-h_0)}{2(\tilde{h}-h_0)} \int_0^1 e^{-(\tilde{h}-h_0)|t-s|} y(s)ds, \quad t \in (0,1) \\ y(1) = 0, y'(0) = 0, \end{cases} \tag{1.8}$$

   where $w = w_0\tilde{h}(\tilde{h} - h_0)^{-1}$.

2. Under the condition $h_0(2\tilde{h} - h_0)(\tilde{h} - h_0)^{-1}(2 + 3(\tilde{h} - h_0)^{-1}) < 1$:

   - there exists a sequence $(\lambda_j)_{j\geq 1}$ of positive real numbers and an orthonormal basis $(\eta_j)_{j\geq 1}$ of $\mathbb{L}^2([0,1])$ such that

$$\lambda_j = \frac{w}{(j\pi - \pi/2)^2} + O(j^{-4})$$

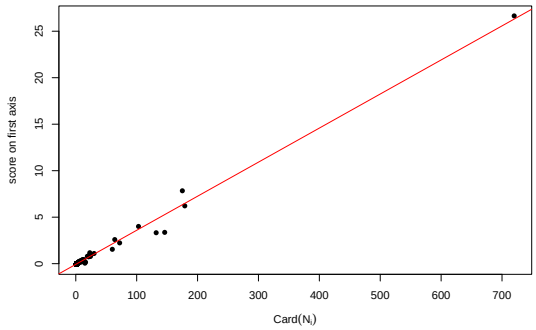

Figure 1.6: Plot of $(\widehat{\xi}_{i1}, \mathrm{Card}(N_i)))_{i=1,\dots,n}$.

$$\sup_{t\in[0,1]} |\eta_j(t) - \sqrt{2}\sin(\pi(2j-1)t/2)| \leq Cj^{-1}.$$

From a numerical point of view, the calculation of the estimator follows the same guidelines as the one explained in the remark of p. 3. The difference lies in the histogram basis which is here adapted to the grid of observation. The method we developed allows us to estimate the eigenfunctions/eigenvalues sequence at rate n^{-1} without regularization nor smoothing. The main key tool for the proof are the inequalities of Bosq (2000) as in [BPRR].

We implement it on a Turkey earthquakes dataset obtained from the website http://www.koeri.boun.edu.tr/sismo/2/earthquake-catalog/ set up by the Kandilli Observatory and Earthquakes Research Institute of the Boğaziçi University. Each point process N_i represents the occurrences of earthquakes at a given location $i \in \{1, \dots, n\}$ ($n = 2063$).

We plot in Figure 1.7 the estimated values of η_j for $j = 1, 2, 3$. We can see that the sign of $\widehat{\eta}_1$ is constant. The first axis is strongly positively correlated with the total number of points $\Pi_i([0,T])$ (with $[0,T]$ the time window of observation): the correlation with the scores on axis 1 $(\widehat{\xi}_{i1})_{i=1,\dots,n}$ and $(\Pi_i([0,T])_{i=1,\dots,n}$ is 0.994. This strong link is illustrated in Figure 1.6.

The value of $\widehat{\lambda}_1 = 2.66\times10^{10}$ represents 97.6% of the variability (meaning that $\widehat{\lambda}_1/(\sum_{j\geq1}\widehat{\lambda}_j) = 0.976$). Despite the important weight of this axis 1, we still choose to analyze axes 2 and 3 and we find a particular structure. First, we remark that $\widehat{\eta}_2$ is approximately constant before July 20, 2017, then negative between July 20, 2017 and October 30, 2020, then positive. Then, from what is explained below we can deduce that, if $\xi_{i,2} > 0$, the process Π_i has a tendency to have less points (compared to the mean) between the July 20, 2017 and October 30, 2020 and more points (still compared to the mean) after October 30, 2020. And, on the contrary, if $\xi_{i,2} < 0$ the process Π_i has a tendency to have more points (compared to the mean) between the July 20, 2017 and October 30, 2020 and less points (still compared to the mean) after October 30, 2020. The third eigenfunction $\widehat{\eta}_3$ has a different behavior, but still shows brutal changes at the same moments in time (corresponding approximately to the 20th of July, 2017 and the 30th of October, 2020). From its signs and variations we can deduce that if $\xi_{i,3} > 0$, the process

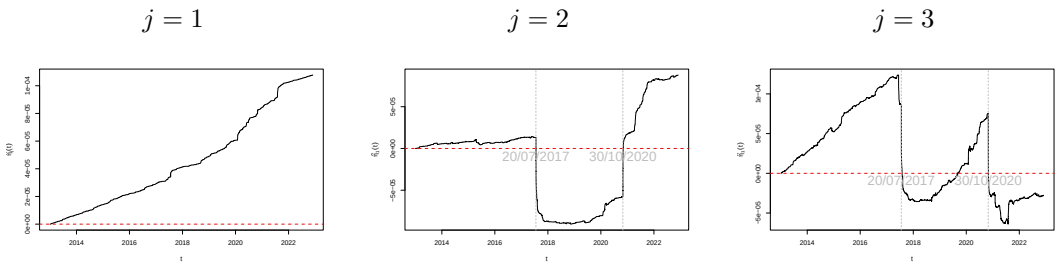

Figure 1.7: Plot of $\widehat{\eta}_j$, for $j = 1, 2, 3$ for the Earthquakes dataset.

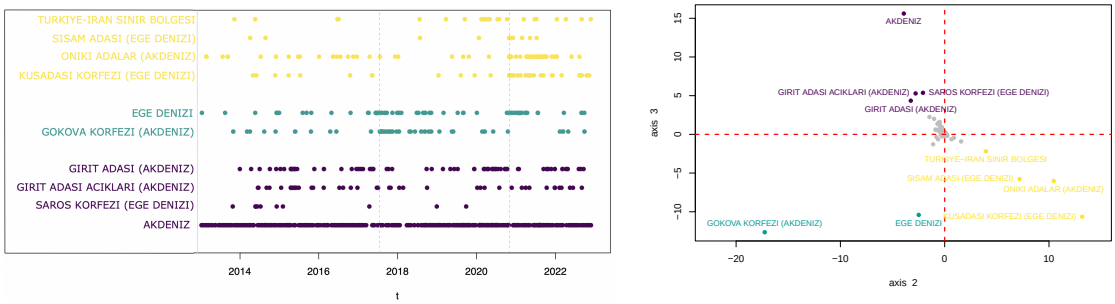

Figure 1.8: Left: Remarkable points processes on axis 2 and 3. Each point represents an Earthquake. Right: PCA scores $(\widehat{\xi}_{i,2}, \widehat{\xi}_{i,3})_{i=1,\dots,n}$ on axis 2 (1.40% or variability) and axis 3 (0.68% of variability).

Π_i has a tendency to be more and more active from the beginning of the time window to July 20, 2017 then less active than the mean then more and more active before October 30, 2020. These two dates corresponds to two major events in the seismic activity of the area: the 2020 Aegean Sea earthquake, with a moment magnitude of 7.0 (the largest magnitude over the period of observation) and the 2017 July 20 6.6 Bodrum–Kos earthquake (Karasözen et al., 2018).

The scores plotted in Figure 1.8 (Left) allow us to detect three groups of locations whose earthquakes have behaviors consistent with our analysis of $\widehat{\eta}_2$ and $\widehat{\eta}_3$ as can be seen on Figure 1.8 (Right).

This approach thus allows us to detect locations with remarkable or atypical behavior in our sample.

1.3 Reduced Order Models [MRRSS]

This project is at the interface between numerical analysis, functional data statistics and Bayesian statistics. We aim at obtaining credible intervals for prediction of infected, recovered or dead people obtained by a Reduced Order Model approach in the work of Bakhta et al. (2021).

The prediction is based on classical simple SIR model where the number of people who are

susceptible of being infected $S(t)$, recovered or dead $R(t)$ and infected $I(t)$ at time t, follows the following dynamics :

$$\begin{aligned} S'(t) &= -\beta(t)I(t)S(t) \\ I'(t) &= -S'(t) - \gamma(t)I(t) \\ R'(t) &= \gamma(t)I(t), \end{aligned} \tag{1.9}$$

with $\beta, \gamma \in \mathbb{L}^\infty([0,T])$ are functional parameters and constant population size $N = S(t) + I(t) + R(t)$. The model with constant coefficients $\beta(t) \equiv \beta$, $\gamma(t) \equiv \gamma$ is a basic epidemiology model that has a lot of variants. The novelty of the approach of Bakhta et al. (2021) is to let the coefficients vary over time that allows the model to fit all possible observations I, R, S as soon as S and I are both differentiable and positive. Indeed, let S and I be differentiable positive functions, then (S, I, R) follows the SIR model (1.9) with coefficients :

$$\beta^*(t) = -\frac{S'(t)}{I(t)S(t)} \text{ and } \gamma^*(t) = -\frac{1}{I(t)}\left(I'(t) - \beta^*(t)I(t)S(t)\right). \tag{1.10}$$

Then there is a correspondance between (S, I, R) and (β, γ) and denote by $S_{\beta,\gamma}$, $I_{\beta,\gamma}$, $R_{\beta,\gamma}$ a solution of (1.9) with coefficients (β, γ).

The problem of interest is the prediction of the future of the series $(I(t), R(t), S(t))$ on an interval $(T, T+\tau)$ with $\tau > 0$ from the observations of $(I(t), R(t), S(t))$ on $[0, T]$. This problem is equivalent to estimating $(\beta(t), \gamma(t))$ on $(T, T+\tau)$ from the observations of $(I(t), R(t), S(t))$ on $[0, T]$.

However, the space $\mathbb{L}^\infty([0, T+\tau))$ is too large and without constraint, the statistical problem of finding $(\beta(t), \gamma(t)) \in \mathbb{L}^\infty([0, T+\tau))^2$ given the observation of $(S(t), I(t), R(t))$ for $t \in [0, T]$ is not identifiable. The method of Bakhta et al. (2021) consists in finding finite dimensional subsets $S_m^\beta = (\varphi_1^\beta, \ldots, \varphi_m^\beta)$ and $S_m^\gamma = (\varphi_1^\gamma, \ldots, \varphi_m^\gamma)$ of $\mathbb{L}^\infty([0, T+\tau])$ and minimizes the least squares contrast

$$(\beta_m^*, \gamma_m^*) \in \arg\min_{(\beta,\gamma) \in S_m^\beta \times S_m^\gamma} \int_0^T \left(I(t) - I_{\beta,\gamma}(t)\right)^2 + \left(R(t) - R_{\beta,\gamma}(t)\right)^2 dt \tag{1.11}$$

on these spaces. We denote by

$$J(\beta, \gamma) = \int_0^T \left(I(t) - I_{\beta,\gamma}(t)\right)^2 + \left(R(t) - R_{\beta,\gamma}(t)\right)^2 dt,$$

and by $J^* = J(\beta_m^*, \gamma_m^*)$ the minimal risk.

The specificity of the Model Order Reduction (MOR) approach lies in the generation of the sets S_m^β and S_m^γ. The aim is that these spaces contains "plausible" values of β and γ. To do that, the space $S_m^\beta \times S_m^\gamma$ is supposed to be an approximation of the manifold of all possible solutions of a detailed compartimental model such as the one described in Di Domenico et al. (2020). In

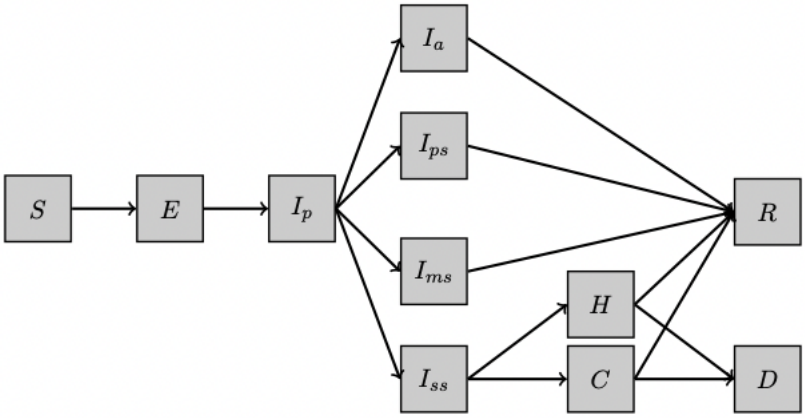

Figure 1.9: `SEI5CHRD` model from Di Domenico et al. (2020)

this model, additional compartments are added compared to the SIR model: E for exposed but non infectious people, I_p for infected and pre-symptomatic people, I_a, I_{ps}, I_{ms}, I_{ss} for infected people with respectively no symptoms (asymptomatic), few symptoms (paucisymptomatics), mild symptoms and severe symptoms, H for hospitalized, C for intensive care unit and D for dead and R for recovered. This model is governed by a system of 11 differential equations and counts 27 parameters $\mu = (\mu_1, \dots, \mu_{27}) \in \mathbb{R}^{27}$. Hence, prediction in this model is quite hard but simulations from this model given a set of parameters $\mu \in \mathbb{R}^{27}$ is feasible. The idea behind the Reduced Basis approach considered in Bakhta et al. (2021) is to simulate a large number of solutions $(S^{(i)}(t), I^{(i)}(t), R^{(i)}(t))$, $t \in [0, T+\tau]$ from a detailed model, then obtain the corresponding values for $(\beta^{(i)}(t), \gamma^{(i)}(t))$ for $t \in [0, T+\tau]$ by solving approximately (1.10) and use classical dimension reduction methods such as fPCA to obtain the spaces S_m^γ and S_m^β. Now the aim of our project is to obtain credible intervals on the predictions.

To do so, we assume that the data

$$\mathbf{y}_{obs} = \{(S(t_h), I(t_h), R(t_h)), h = 0, \dots, p-1\}$$

is generated from the following model:

$$S(t_h) = S_{\beta,\gamma}(t_h) + \varepsilon_h^S,$$

$$I(t_h) = I_{\beta,\gamma}(t_h) + \varepsilon_h^I,$$

$$R(t_h) = n - S(t_h) - I(t_h)$$

with $t_0, \dots, t_{p-1}$ a regular grid of $[0, T]$ and $\{\varepsilon_h^S\}_{h=0,\dots,p-1} \sim_{i.i.d} \mathcal{N}(0, \sigma^2)$, $\{\varepsilon_h^I\}_{h=0,\dots,p-1} \sim_{i.i.d} \mathcal{N}(0, \sigma^2)$, $\sigma > 0$ and $S_{\beta,\gamma}, I_{\beta,\gamma}$ are the solutions of the SIR model equations with parameters

$\beta, \gamma \in \mathbb{L}^\infty([0, +\infty[)$ and n is the (fixed) population size.

We place ourselves in a Bayesian framework and define a sieve prior distribution on β and γ as follows

$$\beta(t) = \sum_{j=1}^{m} b_j \varphi_j^\beta(t) \text{ and } \gamma(t) = \sum_{j=1}^{m} g_j \varphi_j^\gamma(t)$$

where the families of functions $\{\varphi_1^\beta, \ldots, \varphi_m^\beta\}$ and $\{\varphi_1^\gamma, \ldots, \varphi_m^\gamma\}$ are the reduced order basis generated in Bakhta et al. (2021). We assume the basis to be not random (or equivalently we work conditionally on the basis which is supposed to be independent of the data and the prior distribution). Finally, the prior distribution on the coefficients is

$$\mathbf{b} = (b_1, \ldots, b_m) \sim \pi_\beta \text{ and } \mathbf{g} = (g_1, \ldots, g_m) \sim \pi_\gamma.$$

Hence, if the prior distribution (π_β, π_γ) on the coefficients is chosen to be uniform, (β_m^*, γ_m^*) corresponds to the maximum a posteriori (MAP) estimator. This places the approach of Bakhta et al. (2021) in a bayesian framework.

Now the following step is to obtain credible intervals on the prediction of the series $\mathbf{y}_{future} = (S(T+s_h), I(T+s_h), R(T+s_h))_{h=1,\ldots,q}$ with $(s_h)_{h=1,\ldots,q}$ a regular grid of the interval $(0; \tau)$. To do so, we have to approximate the quantiles of the posterior predictive distribution of $\mathbf{y}_{future}$ given $\mathbf{y}_{obs} = (S(t_h), I(t_h), R(t_h))_{h=0,\ldots,p-1}$ that may be written

$$p(\mathbf{y}_{future}|\mathbf{y}_{obs}) = \int_{\mathbb{R}^{2m}} p(\mathbf{y}_{future}|\mathbf{b}, \mathbf{g}) p((\mathbf{b}, \mathbf{g})|\mathbf{y}_{obs}) d\mathbf{b} d\mathbf{g},$$

where $p(\mathbf{y}_{future}|\mathbf{b}, \mathbf{g})$ is the density of the future of the series conditionnally to the parameters $\mathbf{b}$ and $\mathbf{g}$ and $p((\mathbf{b}, \mathbf{g})|\mathbf{y}_{obs})$ is the posterior density. This posterior predictive is intractable since :

- the conditional distribution of $p(\mathbf{y}_{future}|\mathbf{b}, \mathbf{g})$ depends on the function $(\mathbf{b}, \mathbf{g}) \mapsto (S_{\beta,\gamma}, I_{\beta,\gamma})$ which is a fully deterministic, but non linear and non explicit function,

- the posterior density may be written

$$p((\mathbf{b}, \mathbf{g})|\mathbf{y}_{obs}) = \frac{p(\mathbf{y}_{obs}|(\mathbf{b}, \mathbf{g}))\pi_\beta(\mathbf{b})\pi_\gamma(\mathbf{g})}{p(\mathbf{y}_{obs})}$$

  and is also intractable for the same reason ($p(\mathbf{y}_{obs}|(\mathbf{b}, \mathbf{g}))$ depends on the map $(\mathbf{b}, \mathbf{g}) \mapsto (S_{\beta,\gamma}, I_{\beta,\gamma})$).

Then, it is not possible to calculate explicitly $p(\mathbf{y}_{[T;T+\tau)}|\mathbf{y}_{obs})$.

We consider then an Approximate Bayesian Computation approach, that allows us to obtain an approximation of this predictive posterior. The basic idea is to generate a large number of observations from the distribution $(\mathbf{b}^{(1)}, \mathbf{g}^{(1)}), \ldots, (\mathbf{b}^{(N)}, \mathbf{g}^{(N)}) \sim_{i.i.d} \pi_\beta \times \pi_\gamma$ and to keep only

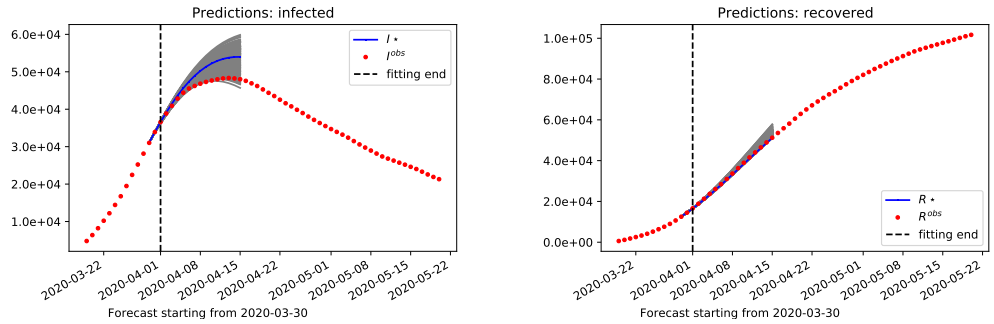

Figure 1.10: Credible intervals (gray), Predictions (plain blue line) of I^* (left) and R^* (right) obtained from the MAP estimators (β_m^*, γ_m^*) and true observations (red dotted line). The black dotted line represents the end of the fitting window T.

the corresponding observations $(S^{(i)}, I^{(i)}, R^{(i)})$, such that

$$J(\beta^{(i)}, \gamma^{(i)}) \leq J^* + \delta,$$

where δ is sufficiently small. The problem is that we are exploring a space which is too large and few observations verify this condition when δ is not too large. Then a second idea is to simulate instead noisy versions $(\beta_{[0,T]}^{(1)}, \gamma_{[0,T]}^{(1)},), \ldots, (\beta_{[0,T]}^{(N)}, \gamma_{[0,T]}^{(N)},)$, by perturbating (β_m^*, γ_m^*):

$$\begin{cases} \beta(t_h) &= \beta_m^*(t_h) + \eta_h^\beta \\ \gamma(t_h) &= \gamma_m^*(t_h) + \eta_h^\gamma \end{cases} \qquad h = 0, \ldots, p-1$$

where $(\eta_0^\beta, \ldots, \eta_{p-1}^\beta)$ and $(\eta_0^\gamma, \ldots, \eta_{p-1}^\gamma)$ is a noise vector that can be correlated (we choose an $AR(1)$ process) and to use importance sampling to obtain estimation of quantiles of the target distribution. In Figure 1.10, we represent the prediction based on the MAP estimators and corresponding credible intervals.

1.4 Perspectives

1.4.1 PCA for point processes: generalizations and applications

The work on the PCA basis for point processes may have several natural extensions. An immediate one is to consider the case of spatio-temporal or spatial point processes which consists in extending the study to points processes on a compact set of $\mathbb{R}^d$. This may allow us, for instance, to take into account the spatial structure of the Earthquakes dataset presented in Section 1.2 but add difficulties linked with the study of multivariate cumulative distributions functions. Another extension, is to consider marked processes.

Other perspectives are opened by the analysis of the PCA scores of the sample. Indeed, as it

is the case for classical PCA, most of the information contained in the sample $N_1, \ldots, N_n$ may be summarized by the matrix of scores $\Xi_J = (\widehat{\xi}_{i,j})_{i=1,\ldots,n;j=1,\ldots,J}$ if J is sufficiently large. Once this score matrix is obtained, it opens the door for the application of classical multivariate methods to the case of point processes, such as regression modelling, discriminant analysis, clustering,... with specificities that remain to explore.

1.4.2 Multilevel PCA for functional imaging data [SR]

This project follows a pilot study (Bazin et al., 2021) that identified both in humans and mouses that gastric inflammation induces modifications of both visible and near-infrared spectra. Modifications of autofluorescence in association with gastric inflammation have also been detected. The identification of gastric inflammatory state is of importance to prevent gastric cancer. This study paves the way toward the development of an optical diagnostic system which would improve on current biopsy-based methods. A research project, led by Professor Dominique Lamarque (Hôpital Ambroise Paré) and Dr Thomas Bazin (Université Versailles Saint-Quentin), which also involves a specialist of polarimetry (Tatiana Novikova, Ecole Polytechnique), was set up with the aim of gathering experimental data on a larger cohort of mice than that resulting from the experiments described in Bazin et al. (2021). Once the data collected, we intend to perform factor analyses to reduce the observations dimensionality and identify the most valuable variables to measure and study. This type of methods is now well-known but the nature of the data will require an expertise, both in probabilities/statistics and in computer sciences. A first idea is to draw inspiration for instance from Zipunnikov et al. (2011) that developed a method to perform functional PCA for high-dimensional multilevel functional data (functional MRI). This method may be adapted to our context and both statistical and computational efficiency of covariance estimation improved using separability ideas from recent works (Masak et al., 2022). After factorial analysis, we aim at setting up classification task based on PCA scores with a supervised learning method (logistic regression for instance).

Chapter 2

Minimax rates in regression models with functional covariates

This chapter is devoted to regression models involving functional data. In all sections (except in Subsection 3.3.1), we consider that we observe a sample

$$\mathcal{D}_n = \{(X_i, Y_i), i = 1, \dots, n\}$$

following the same distribution as a couple (X, Y) of random variables taking values in $\mathcal{X} \times \mathcal{Y}$.

- The space $\mathcal{X}$ is still a Hilbert space, equipped with a scalar product $\langle \cdot, \cdot \rangle_{\mathcal{X}}$ and associated norm $\| \cdot \|_{\mathcal{X}}$. The typical example is $\mathcal{X} = \mathbb{L}^2([0,1])$ with $\langle f, g \rangle_{\mathcal{X}} = \int_0^1 f(t)g(t)dt$, $f, g \in \mathcal{X}$, and $\|f\|_{\mathcal{X}} = \sqrt{\langle f, f \rangle_{\mathcal{X}}}$.
- The space $(\mathcal{Y}, d_{\mathcal{Y}})$ is a metric space. Here, we mainly consider the case where $\mathcal{Y} = \mathbb{R}$ equipped with the distance induced by the absolute value $d_{\mathcal{Y}}(y, y') = |y - y'|$. The case where $\mathcal{Y} = \mathbb{L}^2([0,1])$ with $d_{\mathcal{Y}}(y, y') = \|y - y'\|$ will be considered in subsection 2.1.2.

The aim of the chapter is to give an overview on the minimax rates in this context, assuming different constraints on the relationship between X and Y: linear dependence (section 2.1), single-index constraint (section 2.3.1) or no constraint (section 2.2). We consider, except when otherwise specified, the risk associated to the quadratic loss, that is to say, for an estimator $\widehat{g}$ constructed from $\mathcal{D}_n$,

$$\mathcal{R}_n(\widehat{g}) = \mathbb{E}_P[d^2_{\mathcal{Y}}(Y, \widehat{g}(X))] \tag{2.1}$$

where the expectation $\mathbb{E}_P$ means that the expectation is taken from the distribution $(X, Y) \sim P$. Let $\mathcal{G}$ a class of probability distribution on $\mathcal{X} \times \mathcal{Y}$, the sequence $(r_n)_{n \geq 1}$ is the minimax rate over the class $\mathcal{G}$ if

$$\inf_{\widehat{P}} \sup_{P \in \mathcal{G}} \mathcal{R}_n(\widehat{g}) \asymp r_n,$$

where the infimum is taken over all estimators.

To simplify the mathematical derivations, we assume that both X and Y are centered.

2.1 Functional Linear Regression (FLR) model

The functional linear model has been the object of particular attention during the past decades since it combines an important interest for practitioners and also concentrates interesting theoretical questions. The reference book of Ramsay and Silverman (2010) counts three chapters on the subject. We also refer to Cardot and Sarda (2011) and Hsing and Eubank (2015, Chapter 11).

2.1.1 Linear regression model with scalar output [BR15], [BMR16]

We assume that the relationship between X and Y follows a linear regression model that is to say there exists $\beta^* \in \mathcal{X}$ such that

$$Y = \langle \beta^*, X \rangle_{\mathcal{X}} + \varepsilon, \tag{2.2}$$

where ε is a centered noise, independent of X and with finite variance σ^2. The minimax lower bound for the linear regression model has been proven by Cardot and Johannes (2010). General projection estimators achieving this rate have been defined by Cardot and Johannes (2010) but also by Comte and Johannes (2010, 2012) and in [BR15], [BMR16] (see Section 3.1.1). Ridge estimators have been defined, for instance, by Cai and Yuan (2012) who also prove minimax lower bounds on reproducing kernel Hilbert spaces.

The minimax rates established for the functional linear model depend on regularity assumptions on both β^* and X.

- For X the assumptions are on the covariance operator of X

$$\Gamma : f \in \mathcal{X} \mapsto \mathbb{E}[\langle f, X \rangle_{\mathcal{X}} X].$$

It is assumed the existence of a sequence $\mathbf{v} = (v_j)_{j \geq 1}$ and a constant $d \geq 1$ such that

$$\Gamma \in \mathcal{N}_{\mathbf{v}}(d) = \left\{ \Gamma, d^{-2} \sum_{j \geq 1} v_j^2 \langle f, e_j \rangle_{\mathcal{X}}^2 \leq \|\Gamma f\|_{\mathcal{X}}^2 \leq d^2 \sum_{j \geq 1} v_j^2 \langle f, e_j \rangle_{\mathcal{X}}^2, \text{ for all } f \in \mathcal{X} \right\},$$

for $(e_j)_{j \geq 1}$ an orthonormal basis of $\mathcal{X}$. This assumption is called a link assumption since it characterizes the relationship between the basis $(e_j)_{j \geq 1}$ and the operator Γ.

Choosing the basis $(e_j)_{j \geq 1} = (\psi_j)_{j \geq 1}$ to be the eigenfunctions of the operator Γ, we obtain that

$$\exists d, \Gamma \in \mathcal{N}_{\mathbf{v}}(d) \iff v_j^2 \asymp \lambda_j.$$

Cardot and Johannes (2010) considered two decreasing rates for the sequence $(v_j)_{j \geq 1}$.

	$H_{X,pol}$	$H_{X,exp}$
Lower bound	$n^{-(2\gamma+2b)/(2\gamma+2b+1)}$	$n^{-1}\ln^{2\gamma}(n)$
m^*	$n^{1/(2\gamma+2b+1)}$	$\ln^{2\gamma}(n)$
Upper-bound on $\mathcal{R}_n(\widehat{\beta}_{m^*})$	$n^{-(2\gamma+2b)/(2\gamma+2b+1)}$	$n^{-1}\ln^{2\gamma}(n)$

Table 2.1: Minimax rates of the Functional Linear Model with scalar output on the class $\mathcal{G}^{lin,1}_{\mathbf{v},b}(d,R)$.

$H_{X,pol}$ There exists $\gamma > 1/2$ such that $v_j \asymp j^{-\gamma}$.

$H_{X,exp}$ There exists $\gamma > 0$ such that $v_j \asymp e^{-j^\gamma}$.

Assumption $H_{X,pol}$ is a classical assumption that we can find in most theoretical contributions on FLR. In the case $(e_j)_{j\geq 1} = (\psi_j)_{j\geq 1}$, it is verified by the Brownian motion and Brownian bridge, with $\gamma = 1$, by the fractional Brownian motion, with $\gamma = H + 1/2$ where $H \in]0,1[$ is the Hurst exponent (see Bronski 2003). In the case where $\mathcal{X} = \mathbb{L}^2([0,1])$ $(e_j)_{j\geq 1}$ is the Fourier basis and coincides with $(\psi_j)_{j\geq 1}$ and $\gamma - 1/2$ is an integer, we can characterize $H_{X,pol}$ in terms of the regularity of the operator Γ. Indeed, if $\Gamma \in \mathcal{N}_{\mathbf{v}}(d)$ for a given $d \geq 1$, we have

$$\sum_{j\geq 1} j^{2\gamma}\langle \Gamma f, e_j\rangle^2 = \sum_{j\geq 1} j^{2\gamma}\langle \Gamma f, \psi_j\rangle^2 = \sum_{j\geq 1} j^{2\gamma}\lambda_j\langle f, \psi_j\rangle^2 \lesssim \|f\|^2 < +\infty,$$

then, there exists $\eta > 0$ such that

$$\left|\langle \Gamma f, e_j\rangle^2_{\mathcal{X}}\right| \lesssim j^{-1-2\gamma-\eta}$$

meaning that Γf is at least $\gamma - 1/2$-times differentiable for all $f \in \mathcal{X}$. Then Γ is called finitely smoothing. The case, $H_{X,exp}$ is called infinitely smoothing and corresponds intuitively to the case where Γf is infinitely differentiable. This framework is then related to the functional regularity of the curve X itself via the Karhunen-Loève decomposition.

- For β^* we assume in a similar way that β^* belongs to an ellipsoid of $\mathcal{X}$ i.e. that there exist $b > 0$ and $R > 0$ such that

$$\beta^* \in \mathcal{E}_b(R) = \left\{\beta \in \mathcal{X}, \sum_{j\geq 1} j^{2b}\langle \beta, e_j\rangle^2 \leq R\right\}.$$

  It is also linked with the functional regularity of β^*.

We finally consider the class

$$\begin{aligned}\mathcal{G}^{lin,1}_{\mathbf{v},b}(d,R) = \{&P \text{ dist. of } (X,Y) \text{ s. t. } Y = \langle \beta^*, X\rangle_{\mathcal{X}} + \varepsilon,\\ &\varepsilon \sim \mathcal{N}(0,\sigma^2), \varepsilon \perp X, \beta^* \in \mathcal{E}_b(R), \Gamma \in \mathcal{N}_{\mathbf{v}}(d)\}\end{aligned}$$

The minimax rates for the predictive error (2.1) over $\mathcal{G}^{lin,1}_{\mathbf{v},b}(d,R)$ are given in Table 2.1. The lower bound has been proven by Cardot and Johannes (2010) who defined projection estimators that achieve this rate. In [BR15] and [BMR16] we have proven that the least-squares estimator

$$\widehat{\beta}_m = \arg\min_{\beta\in S_m}\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle\beta, X_i\rangle_{\mathcal{X}})^2\right\}$$

achieves this rate in the case where $S_m = S_m^{PCA} = \text{span}\{\psi_1,\dots,\psi_m\}$ (when the basis $(\psi_j)_{j\geq 1}$ that diagonalizes the covariance operator Γ is known) or, when the ψ_j's are not known, $S_m = \widehat{S}_m^{PCA} = \text{span}\left\{\widehat{\psi}_1,\dots,\widehat{\psi}_m\right\}$ for appropriate choices of the dimension m given in Table 2.1.

2.1.2 Linear Regression Model with functional output [CMR]

In this section, both X and Y are functional variables and we still assume a linear dependency between them i.e. there exists a linear application $B^*:\mathcal{X}\to\mathcal{Y}$ such that

$$Y = B^*X + \varepsilon, \tag{2.3}$$

where ε is centered random variable, independent of X. For sake of simplicity, we assume in the following that $\mathcal{X} = \mathbb{L}^2([0,1])$, $\sigma^2 = \mathbb{E}[\|\varepsilon\|^2_{\mathcal{X}}] < +\infty$ and that there exists an integrable function $\beta^*:[0,1]^2\to\mathbb{R}$ such that

$$B^*f(t) = \int_0^1 \beta^*(s,t)f(s)ds, \qquad f\in\mathbb{L}^2([0,1]).$$

Imaizumi and Kato (2018) proved minimax rates under polynomial assumptions on the rate of decay of the eigenvalues of Γ and on the coefficients of β^* for the $\mathbb{L}^2$ risk. Crambes and Mas (2013) also prove a lower bound based on regularity assumptions on S and an upper-bound based on a regularity assumption on $B^*\Gamma^{1/2}$. We choose in [CMR] a joint assumption on the operator $B^*\Gamma^{1/2}$ and obtain upper and lower bounds for the minimax risk associated to the predictive error (2.1).

In this manuscript, to allow easy comparisons with the FLR with scalar output, the choice has been made to separate both assumptions and assume that

$$B^*\in\mathcal{E}_{b_l,b_r}(R) = \left\{S\in\mathcal{L}_2(\mathcal{X}),\ \sum_{j,k\geq 1} j^{2b_l}k^{2b_r}\langle B\psi_j,\psi_k\rangle^2 \leq R^2\right\},$$

for $b_l, b_r, R > 0$ and consider, as it has been done for the FLR model with scalar output, that $\Gamma\in\mathcal{N}_{\mathbf{v}}(d)$ for a sequence $\mathbf{v}$ satisfying either assumption $H_{X,pol}$ or $H_{X,exp}$ and $d\geq 1$. We refer to [CMR] for the exact formulation of the more general assumption we made on $B^*\Gamma^{1/2}$.

	$H_{X,pol}$	$H_{X,exp}$
Lower bound	$n^{-(2\gamma+2b_l)/(2\gamma+2b_l+1)}$	$n^{-1}\ln^{2\gamma}(n)$
m_l^*	$n^{1/(2\gamma+2b_l+1)}$	$\ln^{2\gamma}(n)$
m_r^*	$+\infty$	$+\infty$
Upper-bound on $\mathcal{R}_n(\widehat{\beta}_{(m_l^*,m_r^*)})$	$n^{-(2\gamma+2b_l)/(2\gamma+2b_l+1)}$	$n^{-1}\ln^{2\gamma}(n)$

Table 2.2: Minimax rates of the Functional Linear Model with functional output on the class $\mathcal{G}^{lin,2}_{\mathbf{v},b_l,b_r}(d,R)$.

The regularity class for the functional linear model with functional output is then

$$\mathcal{G}^{lin,2}_{\mathbf{v},b_l,b_r}(d,R) = \{P \text{ dist. of } (X,Y) \text{ s. t. } Y = B^*X + \varepsilon,$$
$$\varepsilon \text{ Gaussian process}, \varepsilon \perp X, B^* \in \mathcal{E}_{b_l,b_r}(R), \Gamma \in \mathcal{N}_{\mathbf{v}}(d)\}\,.$$

The minimax rates for the predictive error (2.1) over $\mathcal{G}^{lin,1}_{\mathbf{v},b}(d,R)$ are given in Table 2.1. The lower bound can be deduced from [CMR] and the upper-bound is achieved by the least squares estimator

$$\widehat{B}_{m_l,m_r} \in \arg\min_{\beta\in\widehat{S}^{PCA}_{m_l}\otimes\widehat{S}^{PCA}_{m_r}} \frac{1}{n}\sum_{i=1}^{n}\left\|Y_i - \int_0^1 \beta(s,\cdot)X_i(t)dt\right\|^2$$

for appropriate choices of the dimensions m_l^* and m_r^*.

The rates obtained are exactly similar to the ones we had for the FLR with scalar output. Moreover, they do not depend on the "right regularity" of the operator B^*, which seemed to us totally counter-intuitive. Hence the regularization is made only "on the left side" of β meaning that the "right side" is automatically regulated by the integration. It implies, that, from a theoretical viewpoint, predicting a scalar or a function in a linear model has the same statistical complexity. The upper-bound is achieved by projection estimators, as we defined in [CMR] or in Crambes and Mas (2013).

2.1.3 The question of sparsity and the infinite-dimension

In the case where the variable to predict Y depends on d covariates (instead of one in the models studied above), the question of variable selection is posed. In [RLasso] the covariable X is a vector $X = (X^{(1)},\ldots,X^{(d)})$, each $X^{(j)}$ belonging to its own Hilbert space. This kind of data is often referred as multivariate functional data. Two examples are given in figures 2.1 and 2.2. In figure 2.1 the quantity to predict Y is the probability of failure of the nuclear reactor vessel while for the data represented in figure 2.2 it is the mean electric consumption of the day after.

The corresponding multivariate functional linear model can be written as follows:

$$Y = \langle\beta_1^*, X^{(1)}\rangle_{\mathcal{X}_1} + \ldots + \langle\beta_d^*, X^{(d)}\rangle_{\mathcal{X}_d} + \varepsilon, \tag{2.4}$$

with, for all $j = 1,\ldots,p$, $\beta_j^* \in \mathcal{X}_j$ is unknown and ε a centered noise, with finite variance,

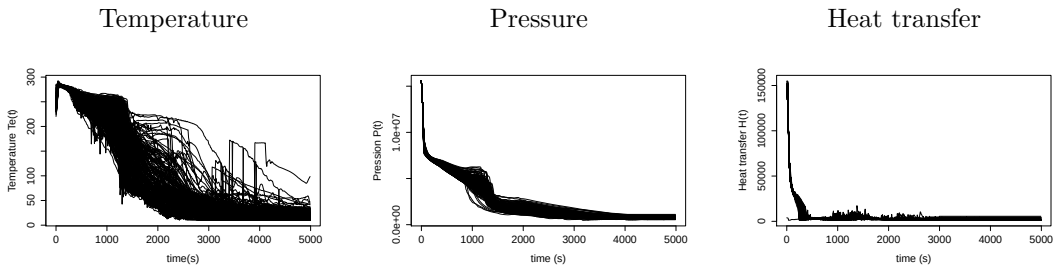

Figure 2.1: Evolution of heat, pressure and heat transfer parameter during simulations of hypothetical loss of coolant accident in a nuclear reactor (see [R18])

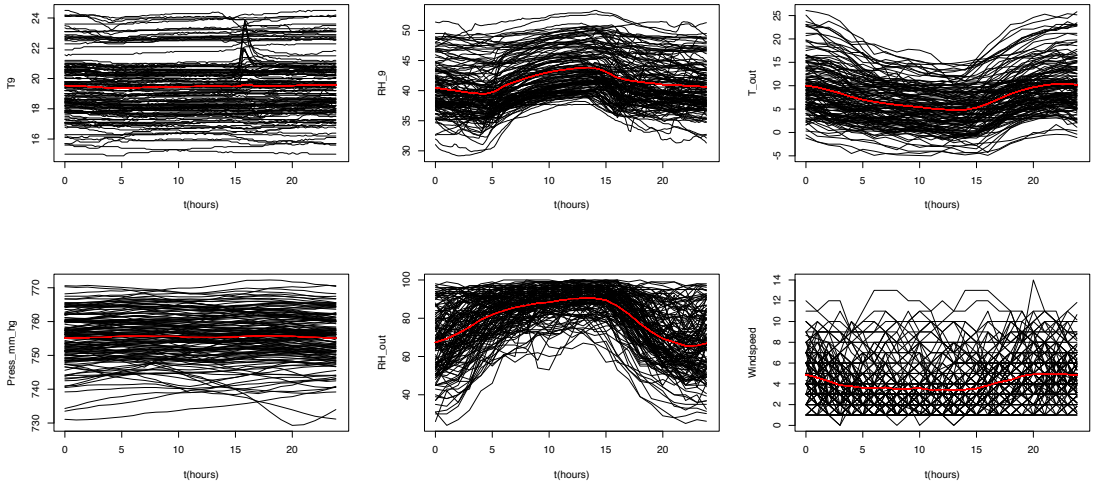

Figure 2.2: Temperature (T9) and humidity (RH_9) of the parent's room, outside temperature (T_out), outside pressure (Press_mm_hg), outside humidity (RH_out) and wind speed during several days (6 among the $d = 24$ covariates considered in [RLasso]).

independent of X.

The product space $\mathcal{X} = \prod_{j=1}^{d} \mathcal{X}_j$ also is a Hilbert space, equipped with its usual scalar product

$$\langle \mathbf{f}, \mathbf{g} \rangle = \sum_{j=1}^{d} \langle f^{(j)}, g^{(j)} \rangle_{\mathcal{X}_j}, \qquad \mathbf{f} = (f^{(1)}, \ldots, f^{(d)}), \mathbf{g} = (g^{(1)}, \ldots, g^{(d)}) \in \mathcal{X},$$

and associated norm $\|\mathbf{f}\|_{\mathcal{X}} = \sqrt{\langle \mathbf{f}, \mathbf{f} \rangle_{\mathcal{X}}}$. The multivariate functional linear model (2.4), is a functional linear model on $\mathcal{X}$,

$$Y = \langle \boldsymbol{\beta}^*, X \rangle_{\mathcal{X}} + \varepsilon,$$

with $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_d^*) \in \mathcal{X}$ to be estimated.

Other interesting cases that are different from what is usually considered as multivariate functional data fall within the scope of model (2.4) :

- the case where the $\mathcal{X}_j$ are Reproducing Kernel Hilbert Spaces. in particular, the question of finding sparse estimators of the vector of coefficient functions is of interest appears in multiple kernel learning problems (Lanckriet et al., 2004; Bach, 2008),
- the case where Y is supposed to depend only on one functional covariate X on $[0, 1]$ and we want to determine if there exist subintervals $(I_j)_j$ of $[0, 1]$ on which β^* is equal to 0. In this case, let $I_1, \ldots, I_d$ d subintervals of $[0, 1]$, $X_j : I_j \to \mathbb{R}$ is the restriction of X to the set I_j. This setting allows us to obtain a slope function that is easy to interpret. In the spirit, it is close to the FLIRTI method (for Functional LInear Regression That is Interpretable) developed by James et al. (2009),
- the case where the data are of different nature (some of the X_j's are functions and the others scalar of vectors for instance).

Let $\mathcal{X}^{(m)}$ a subspace of $\mathcal{X}$ such that $\dim(\mathcal{X}^{(m)}) = m$. We consider the following minimization problem:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m} \in \arg\min_{\boldsymbol{\beta}=(\beta_1,\ldots,\beta_d)\in\mathcal{X}^{(m)}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \langle \boldsymbol{\beta}, X_i \rangle_{\mathcal{X}} \right)^2 + 2 \sum_{j=1}^{d} \lambda_j \|\beta_j\|_{\mathcal{X}_j} \right\}, \tag{2.5}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d) \in]0, +\infty[^d$ is the penalty parameter and $m \in \mathbb{N} \backslash \{0\} \cup \{+\infty\}$. In the finite dimensional case (i.e. $\dim(\mathcal{X}) < +\infty$) this criterion is exactly the same as the one of Lounici et al. (2011).

At the beginning, only the case $m = +\infty$ was considered and the two initial questions that has motivated this work was the following.

1. Is it possible to prove an oracle-type inequality similar to Lounici et al. (2011, Theorem 3.2) in the infinite-dimensional case $\dim(\mathcal{X}) = +\infty$ (which is equivalent to assume that at least one of the $\mathcal{X}_j$'s is infinite-dimensional) ?

2. Is it possible to approach numerically a solution of (2.5) without projecting the data ?

The main difficulty for answering question 1 is that sparsity oracle inequalities are usually obtained under conditions on the design matrix. One of the most common condition is the restricted eigenvalues property. Translated in our context, this assumption may be written as follows.

$(A_{RE(s)})$: There exists a positive number $\kappa = \kappa(s)$ such that

$$\min\left\{\frac{\|\boldsymbol{\delta}\|_n}{\sqrt{\sum_{j\in J}\|\delta_j\|^2_{\mathcal{X}_j}}}, |J|\leq s, \boldsymbol{\delta}=(\delta_1,...,\delta_d)\in\mathcal{X}\backslash\{0\}, \sum_{j\notin J}\lambda_j\|\delta_j\|_{\mathcal{X}_j}\leq c_0\sum_{j\in J}\lambda_j\|\delta_j\|_{\mathcal{X}_j}\right\}\geq\kappa,$$

with $\|f\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^n\langle f,\mathbf{X}_i\rangle^2_{\mathcal{X}}}$ the empirical norm on $\mathcal{X}$ naturally associated with our problem.

The constant κ has to be strictly positive since its inverse appears in the upper-bound. For instance, applying Lounici et al. (2011, Theorem 3.2) with our notations gives us, for the case $\dim(\mathcal{X})<+\infty$ and the noise ε is Gaussian, for λ_j sufficiently large,

$$\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}-\boldsymbol{\beta}^*\|_n^2\leq\min_{\boldsymbol{\beta}\in\mathcal{X},J(\boldsymbol{\beta})\leq s}\left\{\frac{96}{\kappa^2}\sum_{j\in J(\boldsymbol{\beta})}\lambda_j^2+2\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_n^2\right\},$$

with probability larger than $1-2d^{1-q}$ with $J(\boldsymbol{\beta})=\mathrm{Card}\{\mathrm{j},\beta_{\mathrm{j}}\neq 0\}$.

The next lemma, proved in [RLasso], shows that assumptions like $(A_{RE(s)})$ can not hold when the dimension of the space is too large. We first introduce new notations: let, for $J\subset\{1,\dots,p\}$,

$$\mathcal{X}_J=\prod_{j\in J}\mathcal{X}_j \text{ and } \widehat{\Gamma}_J\mathbf{f}=(f_j)_{j\in J}\in\mathcal{X}_J\to\left(\frac{1}{n}\sum_{i=1}^n\sum_{j\in J}\langle f_j,X_i^{(j)}\rangle_{\mathcal{X}_j}X_i^{(j')}\right)_{j'\in J},$$

the product space of the $\mathcal{X}_j$ for $j\in J$ and the empirical covariance operator of the data $(X_i^{(j)}, i=1,\dots,n; j\in J)$.

Lemma 6 ([RLasso]). Suppose that there exists $J\subset\{1,\dots,p\}$ such that $\dim(\mathcal{X}_J)>\mathrm{rk}(\widehat{\Gamma}_J)$, then, for all $s\geq|J|$, for all $c_0>0$

$$\min\left\{\frac{\|\boldsymbol{\delta}\|_n}{\sqrt{\sum_{j\in J}\|\delta_j\|^2_{\mathcal{X}_j}}}, |J|\leq s, \boldsymbol{\delta}=(\delta_1,...,\delta_d)\in\mathcal{X}\backslash\{0\}, \sum_{j\notin J}\lambda_j\|\delta_j\|_{\mathcal{X}_j}\leq c_0\sum_{j\in J}\lambda_j\|\delta_j\|_{\mathcal{X}_j}\right\}=0.$$

Idea of proof. $\widehat{\Gamma}_J$ is a linear operator on $\mathcal{X}_J$. If $\dim(\mathcal{X}_J)>\mathrm{rk}(\widehat{\Gamma}_J)$, this means that $\dim(\mathrm{Ker}(\widehat{\Gamma}_J))\geq 1$ and then we can construct $\boldsymbol{\delta}_J=(\delta_j)_{j\in J}\in\mathcal{X}_J\backslash\{0\}$ such that $\widehat{\Gamma}_J\boldsymbol{\delta}_J=0$. Defining now $\delta_j=0$ for $j\notin J$, we define an element $\boldsymbol{\delta}\in\mathcal{X}$ satisfying the constraint $\sum_{j\notin J}\lambda_j\|\delta_j\|_{\mathcal{X}_j}\leq c_0\sum_{j\in J}\lambda_j\|\delta_j\|_{\mathcal{X}_j}$ and such that $\|\boldsymbol{\delta}\|_n^2=\|\widehat{\Gamma}_J\boldsymbol{\delta}_J\|^2_{\mathcal{X}_J}=0$. □

Since $\mathrm{rk}(\widehat{\Gamma}_J)\leq n$, the conclusion that can be drawn from Lemma 6 is that no Restricted Eigenvalues assumptions can hold as soon as the ambient space $\mathcal{X}$ is infinite-dimensional (and

even when it is finite-dimensional, as soon as the dimension is larger than the number of observations). However, restrictions to finite-dimensional spaces may be considered. Let for $m \geq 1$ and $s \in \{1, \dots, p\}$, $\mathcal{X}^{(m)}$ be a m-dimensional subspace of $\mathcal{X}$,

$$\tilde{\kappa}_n^{(m)}(s) \quad := \\ \min \left\{ \frac{\|\boldsymbol{\delta}\|_n}{\sqrt{\sum_{j\in J} \|\delta_j\|_{\mathcal{X}_j}^2}}, |J| \leq s, \boldsymbol{\delta} = (\delta_1, ..., \delta_d) \in \mathcal{X}^{(m)} \backslash \{0\}, \sum_{j \notin J} \lambda_j \|\delta_j\|_{\mathcal{X}_j} \leq 3 \sum_{j \in J} \lambda_j \|\delta_j\|_{\mathcal{X}_j} \right\}. \tag{2.6}$$

it is reasonable to assume that, for m not too large, $\tilde{\kappa}_n^{(m)}(s) > 0$. For s fixed, the sequence $(\tilde{\kappa}_n^{(m)}(s))_{m\geq 1}$ is a non-increasing sequence, equal to 0 up to a certain rank, which seems to be related, in a complicated way, to the decreasing rates of the eigenvalues of the operators $\{\widehat{\Gamma}_J, |J| \leq s\}$ which is related to the regularity of the data (see Section 2.1.1 and also the examples given in Section 2.3 of [RLasso]).

The following proposition is proved in [RLasso] under a unique assumption of subgaussianity of the noise ε and under some constraint on the basis that is detailed in [RLasso]. The result is written here in the case of Gaussian noise for simplicity.

Proposition 7 ([RLasso]). Let $q > 0$ be fixed and choose

$$\lambda_j = r_n \left(\frac{1}{n} \sum_{i=1}^n \|X_i^j\|_{\mathcal{X}_j}^2 \right)^{1/2} \text{ with } r_n = A\sigma \sqrt{\frac{q \ln(d) + \ln(2)}{n}} \quad (A \geq 4). \tag{2.7}$$

With probability larger than $1 - d^{1-q}$, for all $m \geq 1$,

$$\left\| \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m} - \boldsymbol{\beta}^* \right\|_n^2 \leq \min_{\boldsymbol{\beta} \in \mathcal{X}^{(m)}, |J(\boldsymbol{\beta})| \leq s} \left\{ \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_n^2 + \frac{9}{4(\tilde{\kappa}_n^{(m)})^2} \sum_{j \in J(\boldsymbol{\beta})} \lambda_j^2 \right\} \tag{2.8}$$

using the convention $1/0 = +\infty$ in the case where $\tilde{\kappa}_n^{(m)} = 0$.

The upper-bound is the sum of an approximation term $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_n$ and a term related to both the penalty and the quantity $\tilde{\kappa}_n^{(m)}$. The proof combine ideas from the proofs of Lounici et al. (2011, Theorem 3.2) and Bellec and Tsybakov (2017, Proposition 5). In particular, the constants appearing in the upper-bound are better than those of Lounici et al. (2011, Theorem 3.2) (2 replaced by one for the "bias term", which implies that the oracle-inquality is "sharp", and 96 replaced by 9/4 in the term associated to the penalty).

We also remark that the two terms in Eq. (2.8) can be controlled to obtain an upper-bound on the convergence rates. For this, we need additional moment assumptions on X and consider, as in sections 2.1.1 and 2.1.2, the regularity classes $\mathcal{N}_{\mathbf{v}}(d)$, for the covariance operator Γ, with $v_j \asymp j^{\gamma}$ (assumption $H_{X,pol}$), and the regularity class $\mathcal{E}_b(R)$ for $\boldsymbol{\beta}^*$. We also define a theoretical version $\kappa^{(m)}(s)$ of $\tilde{\kappa}_n^{(m)}(s)$ (replacing in the definition the empirical norm $\|\cdot\|_n = \|\widehat{\Gamma}^{1/2} \cdot \|$ by its

theoretical counterpart $\|\cdot\|_\Gamma = \|\Gamma^{1/2}\cdot\|$), and suppose that

$$\kappa^{(m)}(s) \asymp m^{-\gamma(s)},$$

for $\gamma(s) \geq 1/2$, $s \in \{1,\dots,p\}$.

We choose, for all $j = 1,\dots,p$,

$$\lambda_j = A\sigma\sqrt{\frac{\ln(n)+\ln(d)}{n}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|X_i^{(j)}\|_j^2},$$

with $A > 0$ a numerical constant and

$$m_n^* \asymp \left(\frac{n}{s(\ln(n)+\ln(d))+\ln^2(n)}\right)^{\frac{1}{2b+2\gamma(s)+2\gamma}}.$$

Under these assumptions, and assuming in addition $\boldsymbol{\beta}^*$ has less than s non null coefficients (i.e. $|J(\boldsymbol{\beta}^*)| \leq s$) there exist two constants $C, C' > 0$ such that, with probability larger than $1 - C/n$

$$\sup_{\boldsymbol{\beta}^*\in\mathcal{E}_b(R),\boldsymbol{\Gamma}\in\mathcal{N}_v(c)} \left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m_n^*} - \boldsymbol{\beta}^*\right\|_\Gamma^2 \leq C'\left(\frac{s(\ln(d)+\ln(n))+\ln^2(n)}{n}\right)^{\frac{b+\gamma}{b+\gamma(s)+\gamma}}. \tag{2.9}$$

Comparing with the minimax rates in the functional linear model given in Table 2.1, the rate we obtain is probably not minimax optimal. Since the bias term has the same order than the one of the functional linear model, the problem comes certainly with the order of the term related to the penalty which is

$$(\kappa^{(m)}(s))^{-2}\sum_{j\in J(\boldsymbol{\beta}^*)}\lambda_j^2 \asymp m^{2\gamma(s)}s\frac{\ln(n)+\ln(d)}{n}$$

whereas an order of m/n (up to eventual $\ln(n)$, $\ln(d)$ and s) is expected. A possible solution to solve this problem could be to change the penalty in the criterion (2.5) from e.g. an adaptive one which is a perspective of this work.

Question 2. has been solved by using a coordinate descent algorithm in the spirit of the glmnet algorithm (Friedman et al., 2010). The difficulty is that the algorithm is based on the so-called group-wise orthonormality criterion which, translated to our context, is equivalent to suppose that the operators $\widehat{\Gamma}_j$ are all equal to the identity, up to a multiplicative constant, which is impossible when $\dim(\mathcal{X}_j) = +\infty$ since $\widehat{\Gamma}_j$ is not invertible. To overcome this problem, the Groupwise-Majorization-Descent algorithm developed in finite-dimensional contexts by Yang and Zou (2015) can be adapted in the infinite-dimensional context (see sections 5 and 6 of [RLasso]).

We choose λ_j as described above and remark that λ_j is entirely defined by the quantity A and we denote by $\boldsymbol{\lambda}(A)$ the corresponding vector of values of $\lambda_1,\dots,\lambda_d$. Then we define

	Model 1			Model 2		
	$\widehat{\boldsymbol{\lambda}}^{(CV)}$	$\widehat{\boldsymbol{\lambda}}^{(\widehat{\sigma}^2)}$	$\widehat{\boldsymbol{\lambda}}^{(BIC)}$	$\widehat{\boldsymbol{\lambda}}^{(CV)}$	$\widehat{\boldsymbol{\lambda}}^{(\widehat{\sigma}^2)}$	$\widehat{\boldsymbol{\lambda}}^{(BIC)}$
Support recovery of $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\lambda}},\infty}$ (%)	0	100	0	2	100	4
Support recovery of $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\lambda}},\widehat{m}}$ (%)	/	100	/	/	100	/

Table 2.3: Percentage of times where the true support has been recovered among 50 Monte-Carlo replications of the LASSO estimator.

the grid $\mathbf{A} = \{A_1, \dots, A_{\max}\}$ of possible values for A such that $A_1 < \dots < A_{\max}$ and that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}(A_{\max}),m} = (0, \dots, 0)$. The algorithm defined in [RLasso] is the following.

Algorithm 1 Coordinate descent algorithm inspired from Yang and Zou (2015) to approach $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}$

For all $A_r \in \mathbf{A}$: initialize $\widehat{\boldsymbol{\beta}}^{(0)}_{\boldsymbol{\lambda}(A_r),m} = \widehat{\boldsymbol{\beta}}^{(final)}_{\boldsymbol{\lambda}(A_{r-1}),m}$.
repeat
  For $j = 1, \dots, d$,
$$\widehat{\beta}_j^{(\ell)} = \arg\min_{\beta \in \mathbb{R}} \tilde{\gamma}_{n,j}(\beta).$$
until $\frac{1}{n}\sum_{i=1}^n \langle \widehat{\boldsymbol{\beta}}^{(\ell)}_{\boldsymbol{\lambda},m} - \widehat{\boldsymbol{\beta}}^{(\ell-1)}_{\boldsymbol{\lambda},m}, \mathbf{X}_i\rangle^2_{\mathcal{X}} \leq s$ or maximal number of iterations reached
Return $\widehat{\boldsymbol{\beta}}^{(final)}_{\boldsymbol{\lambda}(A_r),m}$.

Here $\tilde{\gamma}_{n,j}(\beta_j)$ is a majorant of the target quantity

$$\gamma_{n,j}(\beta_j) = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle_{\mathcal{X}}\right)^2 + 2\lambda_j \|\beta_j\|_{\mathcal{X}_j},$$

obtained by fixing the coordinates of $\boldsymbol{\beta}$ except the j-th one. The idea of initializing $\widehat{\boldsymbol{\beta}}^{(0)}_{\boldsymbol{\lambda}(A_r),m}$ with $\widehat{\boldsymbol{\beta}}^{(final)}_{\boldsymbol{\lambda}(A_{r-1}),m}$ comes from Friedman et al. (2010) and, according to the authors, leads to a more stable and faster algorithm.

On simulated data, we are able to provide methods to select $\boldsymbol{\lambda}$ and the couple $(\boldsymbol{\lambda}, m)$, that have good practical support recovery properties for $\boldsymbol{\beta}^*$ (we refer to Section 6 of [RLasso] for the description of the two models considered to simulate the data and to section 5.2 for the description of the three methods to select $\boldsymbol{\lambda}$). The dimension $\widehat{m}$ is selected by minimizing a penalized contrast criterion

$$\widehat{m} \in \arg\min_m \left\{ \frac{1}{n}\sum_{i=1}^n \left(Y_i - \langle \widehat{\boldsymbol{\beta}}_{\widehat{\lambda},m}, \mathbf{X}_i\rangle_{\mathcal{X}}\right)^2 + \kappa\sigma^2 \frac{m\log(n)}{n} \right\}.$$

The explication for using the criterion above will be given in subsection 3.1.1. However, despite the good support recovery properties, we observe that both estimators $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\lambda}},\infty}$ and $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\lambda}},\widehat{m}}$ are biased, which is in coherence with what is also observed for Lasso estimators in the finite-dimensional case (see Giraud 2015, Section 4.2.5). This bias is corrected with a Ridge estimator on the Lasso

Example 1 Example 2

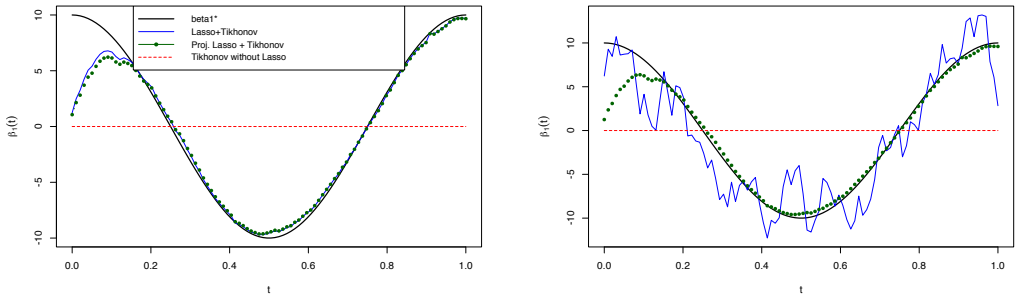

Figure 2.3: Plot of β_1^* (solid black line), the solution of the Tikhonov regularization on the support of the Lasso estimator (dashed blue line) and on the whole support (dotted red line).

	Lasso + Tikhonov	Proj. Lasso + Tikhonov	Tikhonov without Lasso
Example 1	7.5 min	9.3 min	36.0 min
Example 2	7.1 min	16.6 min	36.1 min

Table 2.4: Computation time of the estimators.

support calculated - without projecting the data - by the following stochastic gradient descent algorithm initialized at the estimator $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\lambda}},\infty}$ or $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\lambda}},\widehat{m}}$ (depending on context).

Algorithm 2 Stochastic gradient descent algorithm on the Lasso support

Initialize $\widetilde{\boldsymbol{\beta}}^{(0)} = \widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\lambda}},\widehat{m}}$ or $\widetilde{\boldsymbol{\beta}}^{(0)} = \widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\lambda}},\infty}$.
repeat

$$\tilde{\boldsymbol{\beta}}^{(\ell)} = \tilde{\boldsymbol{\beta}}^{(\ell-1)} - \alpha_\ell \, \nabla_{\tilde{\boldsymbol{\beta}}^{(\ell-1)}} \gamma_{n,\rho}^{(R)},$$

where $\nabla_f \gamma_{n,\rho}^{(R)}$ is the gradient of the Ridge criterion $\gamma_{n,\rho}^{(R)}$ with parameter ρ at the point f and $\alpha_\ell = \alpha_1 \ell^{-1}$.
until $\frac{1}{n}\sum_{i=1}^n \langle \tilde{\boldsymbol{\beta}}^{(\ell)} - \tilde{\boldsymbol{\beta}}^{(\ell-1)}, \mathbf{X}_i \rangle_{\mathcal{X}}^2 \leq s$ or maximal number of iterations reached

The parameter ρ is selected by cross-validation.

We see in Figure 2.3 that the resulting estimators perform well, especially compared to the Tikhonov regularization without Lasso (i.e. with all the $d = 7$ covariates). The computation times also are in favor of the Lasso + Tikhonov estimation procedure.

2.2 "Nonparametric" regression model [CR14], [CR16]

Another model that has been widely studied in the literature is the so-called "non parametric" regression model (see Chapter 5 of Ferraty and Vieu 2006). For observations following this

model, we de not assume a particular form of the regression functional $r(x) = \mathbb{E}[Y|X = x]$. In this context we studied in [CR14] the minimax rates for the estimation of the cumulative distribution function (c.d.f.) of Y given X

$$F^x(y) = \mathbb{P}(Y \leq y|X = x),$$

and in [CR16] the minimax rates for the estimation of the regression function

$$r(x) = \mathbb{E}[Y|X = x].$$

We started from the articles of Ferraty et al. (2006); Ferraty and Vieu (2000) and consider kernel estimators of the form

$$\widehat{F}^x_{h,\mathfrak{d}}(y) = \sum_{i=1}^n W^{(i)}_{h,\mathfrak{d}}(x)\mathbf{1}_{\{Y_i \leq y\}} \text{ and } \widehat{m}^x_{h,\mathfrak{d}} = \sum_{i=1}^n W^{(i)}_{h,\mathfrak{d}}(x)Y_i$$

for the c.d.f. and regression function respectively, where the weights are defined as follows

$$W^{(i)}_{h,\mathfrak{d}}(x) = \frac{K_h(\mathfrak{d}(X_i, x))}{\sum_{j=1}^n K_h(\mathfrak{d}(X_j, x))}$$

with $K_h(\cdot) = h^{-1}K(\cdot/h)$, K is a kernel function, $h > 0$ the bandwidth and $\mathfrak{d}$ a pseudo-distance on $\mathcal{X}$. Assuming that

$$\left|F^x(y) - F^{x'}(y)\right| \leq L\mathfrak{d}^b(x, x'), \tag{2.10}$$

for a constant $L > 0$ and $b \in]0, 1]$, Ferraty et al. (2006) obtained convergence rates of the form

$$\sup_{y\in S}\left|\widehat{F}^x_{h,\mathfrak{d}}(y) - F^x(y)\right| = O\left(h^b + \sqrt{\frac{\ln(n)}{n\varphi_{x,\mathfrak{d}}(h)}}\right) \qquad a.s.^1$$

with $\varphi_{x,\mathfrak{d}}(h) = \mathbb{P}(\mathfrak{d}(X, x) \leq h)$ is the small ball probability of X associated to the pseudo-distance $\mathfrak{d}$ and S a compact subset of $\mathcal{X}$. The final rates then depends on the asymptotic behavior of the small ball probability when $h \to 0$. Two main cases then occur.

- The small-ball probability is polynomial in h i.e. $\varphi_{x,\mathfrak{d}}(h) \sim_{h\to 0} h^{2\gamma}$ for $\gamma > 0$. In that case, choosing $h \sim (\log(n)/n)^{1/(2b+2\gamma)}$ they obtain a polynomial rate of order:

$$\sup_{y\in S}\left|\widehat{F}^x_{h,\mathfrak{d}}(y) - F^x(y)\right| = O\left((\log(n)/n)^{b/(2b+2\gamma)}\right) \qquad a.s. \tag{2.11}$$

  However, we have to be careful with this rate:

  – If $\mathfrak{d}$ is a distance (and not a pseudo-distance) and $\mathrm{rk}(\Gamma) = +\infty$, these polynomial rates

[1] the notion considered by Ferraty and Vieu (2000) is the notion of almost complete convergence that imply both a.s. convergence and convergence in probability.

are impossible (see Mas 2012; Azaïs and Fort 2013). Then, in the case $\mathrm{rk}(\Gamma) = +\infty$, $\mathfrak{d}$ must be a pseudo-distance. For instance, the case

$$\mathfrak{d}_m(x, x') = \sqrt{\sum_{j=1}^{m} \langle x - x', e_j \rangle_{\mathcal{X}}^2},$$

with $m \in \mathbb{N}^*$ and $\{e_1, \ldots, e_m\}$ an orthonormal family of $\mathcal{X}$ gives a small-ball probability of order

$$\varphi_{x,\mathfrak{d}}(h) \asymp h^m.$$

– Since the pseudo-distance $\mathfrak{d}$ appears in the estimation procedure, it is chosen by the statistician. Hence assumption (2.10) seemed to us too strong in the case of projection pseudo-distance. Indeed, it implies that the conditional distribution of Y given X only depends on the coefficients $(\langle X, e_1 \rangle_{\mathcal{X}}, \ldots, \langle X, e_m \rangle_{\mathcal{X}})$ and that the family $\{e_1, \ldots, e_m\}$ is known (since they appear in the definition of the estimator).

To overcome these difficulties, we adopted in [CR14] the following assumption:

$$\left| F^x(y) - F^{x'}(y) \right| \leq L \|x - x'\|_{\mathcal{X}}^b, \qquad x, x' \in \mathcal{X}, \tag{2.12}$$

for a constant $L > 0$ and $b \in]0, 1]$. We obtained the following results on the pointwise risk (predictive version under additional assumptions are also proved in [CR14]).

Proposition 8 ([CR14]). Under some assumptions on the distribution of the coefficients $(\langle X, e_j \rangle_{\mathcal{X}})$:

$$\mathbb{E}\left[\left\| \widehat{F}_{h,\mathfrak{d}_m}^{x_0} - F^{x_0} \right\|^2 \right] \leq C \left(h^{2b} + \left(\sum_{j>m} \mathrm{Var}(\langle X, e_j \rangle_{\mathcal{X}}^2 \right)^b + \left(\sum_{j>m} \langle x_0, e_j \rangle_{\mathcal{X}}^2 \right)^b + \frac{1}{n \varphi_{x_0,\mathfrak{d}_m}(h)} \right). \tag{2.13}$$

The two additional terms compared to (2.11) are due to the fact that the distance $\mathfrak{d}_m$ appearing in the kernel differs now from the one appearing in assumption (2.12). They both decrease to 0 when $m \to \infty$ but not sufficiently fast to achieve a polynomial rate similar to the one of Ferraty and Vieu (2000) (in addition it can be shown the small ball probability $\varphi_{x_0,\mathfrak{d}_m}(h)$ is of order h^m, which degrades the variance term when m is large). On the contrary, the estimator $\widehat{F}_{h,\infty}^{x_0}$ obtained with $\mathfrak{d}_\infty(x, x') = \|x - x'\|$ have the following bias-variance decomposition

$$\mathbb{E}\left[\left\| \widehat{F}_{h,\infty}^{x_0} - F^{x_0} \right\|^2 \right] \leq C \left(h^{2b} + \frac{1}{n \varphi_{x_0,\infty}(h)} \right), \tag{2.14}$$

but the behavior of the small ball probability $\varphi_{x_0,\infty}(h) = \mathbb{P}(\|x - x_0\| \leq h)$ associated to the distance $\mathfrak{d}_\infty(x, x') = \|x - x'\|$ does not allow for polynomial convergence rates in the case where $\mathrm{rk}(\Gamma) = +\infty$. The conclusion that can be drawn is that, with the upper-bound (2.13), non

Assumption	$H_{X,L}$	$H_{X,M}$	$H_{X,F}$
Lower bound	$(\ln(n))^{-2b/\gamma'}$	$\exp\left(-\frac{2b}{c_1^{1/\gamma'}}\ln^{1/\gamma'}(n)\right)$	$n^{-\frac{2b}{2b+d}}$
Order of h^*	$\ln^{-1/\gamma'}(n)$	$\exp\left(-\frac{1}{c_1}\ln^{1/\gamma'}(n)\right)$	$n^{\frac{1}{2b+d}}$
$\mathcal{R}_n(\widehat{F}_{h^*})$	$(\ln(n))^{-2b/\gamma'}$	$\exp\left(-\frac{2b}{c_1^{1/\gamma'}}\ln^{1/\gamma'}(n)\right)$	$n^{-\frac{2b}{2b+d}}$

Table 2.5: Minimax rates for the estimation of the conditional c.d.f. [CR14] and the regression function [CR16] in the "non parametric" model.

parametric convergence rates of the form (2.11) are not achievable. The open question at this step was if another estimator could achieve non parametric convergence rates under assumption (2.12). To answer the question we proved lower bounds under some assumptions on the behaviors of small ball probability functions described below.

$H_{X,L}$ There exist $\gamma' > 0$ and $c_1 \in \mathbb{R}$, $c_2 > 0$ such that

$$\varphi_{x_0,\infty}(h) \asymp h^{c_1}\exp(-c_2 h^{-\gamma'}).$$

$H_{X,M}$ There exist $\gamma' > 1$ and $c_1, c_2 \in \mathbb{R}$, such that

$$\varphi_{x_0,\infty}(h) \asymp h^{c_1}\exp(-c_2\ln^{\gamma'}(1/h)).$$

$H_{X,F}$ There exists a constant $d > 0$, such that $\varphi_{x_0,\infty}(h) \asymp h^d$.

The assumptions $H_{X,L}$ and $H_{X,M}$ are related to the assumptions $H_{X,pol}$ and $H_{X,exp}$ made in subsection 2.1. For instance, if X is a Gaussian process and satisfies $H_{X,pol}$, then Assumption $H_{X,L}$ is satisfied with $c_1 = (3-\gamma)/(2\gamma-1)$, $c_2 = \gamma(2\gamma/(2\gamma-1))^{1/(2\gamma-1)}$ and $\gamma' = 1/(\gamma-1/2)$ (Hoffmann-Jørgensen et al. 1979, Theorem 4.4 and example 4.5, p.333-334). The second case $H_{X,M}$ typically happens when the eigenvalues of the covariance operator decrease exponentially fast (see Dunker et al. 1998, Proposition 4.3 p.12). In the case where $c\exp(-2j)/j \leq \lambda_j \leq C\exp(-2j)/j$, we have $c_2 = 1/2$ and $\gamma' = 2$. (Hoffmann-Jørgensen et al. 1979, Theorem 4.4 and example 4.7, pp. 333 and 336).

The obtained rates are given in Table 2.5. The corresponding lower bounds are proved in Section 4.3 [CR14] and are achieved by $\widehat{F}_{x,\infty}(h^*)$ for particular choices of h^*. We obtain similar rates for the estimation of the regression function in [CR16] for the risk associated to the pointwise mean squared error. The proof of the lower bound relies on the general scheme described in Tsybakov (2009) with inspiration from a similar bound proven in regression for functional data by Mas (2012) and in c.d.f. for univariate data by Brunel et al. (2010).

2.3 Perspectives

2.3.1 Single and multiple index models [DRR]

Single-index models are intermediate models between the linear model, which may be too restrictive for some applications, and the "non parametric" model, which suffers from the curse of dimensionality. They have been defined first for classical multivariate statistics (see e.g. Härdle et al. 1997) and extended naturally to the case of functional data by Ferraty et al. (2011). The model can be written

$$Y = g^*(\langle \beta^*, X \rangle) + \varepsilon,$$

where both the link function $g^* : \mathbb{R} \to \mathbb{R}$ and the index parameter β^* are unknown.

The main difficulties for estimation in this model come from the nonlinearity of g^* and the fact that the support of g^* can not be naturally considered as a compact subset of $\mathbb{R}$. To overcome these difficulties, we consider a Bayesian approach based on hybrid-location scale mixtures of normal prior, introduced by Naulet and Rousseau (2017). We obtain posterior concentration rates that are the maximum between the minimax rate of estimation for the functional linear model given in Table 2.1 and the minimax rate of the estimation of g^* if β^* is known. We have obtained posterior concentration rate in empirical norm

$$d_n((\beta, g), (\beta^*, g^*)) = \left(\frac{1}{n} \sum_{i=1}^{n} \left(g(\langle \beta, X_i \rangle) - g^*(\langle \beta^*, X_i \rangle) \right)^2 \right)^{1/2}.$$

Theorem 9. Suppose that there exists $p > 2$ such that $\sup_{j \geq 1} \mathbb{E}[\langle X, \psi_j \rangle^p / \lambda_j^{p/2}] < +\infty$, that $H_{X,pol}$ is verified and that there exists $\alpha, L > 0$ such that

$$\|g^*\|_\alpha \leq L \qquad \text{and} \qquad \mathbb{E}[g^*(\langle \beta^*, X \rangle)] < +\infty$$

and, $b, R > 0$ such that $\beta^* \in \mathcal{E}_b(R)$. Then, under some assumptions on the prior distribution Π of (β, g),

$$\mathbb{E}\left[\Pi\left(d_n((\beta, g), (\beta^*, g^*)) \gtrsim r_n^{up} \right) | \mathcal{D}_n \right] \xrightarrow[n \to +\infty]{} 0,$$

with

$$r_n^{up} = \max\left\{ n^{-\min\{\alpha;1\}(b+\gamma)/(2(b+\gamma)+1)}; n^{-\alpha/(2\alpha+1)} \right\}.$$

The proof of the lower bound is in progress but we already have obtained the following partial result.

Lemma 10. Suppose $H_{X,pol}$ is verified. There exists a positive quantity $c > 0$ and an integer n^* (depending only on the sequence $(\lambda_j)_{j \geq 1}$, b, α, L and R) such that, for all $n \geq n^*$,

$$\inf_{(\widehat{\beta}, \widehat{g})} \sup_{\beta^* \in \mathcal{E}_b(R), \|g^*\|_\alpha \leq L} \mathbb{E}[d_n((\widehat{\beta}, \widehat{g}), (\beta^*, g^*))] \geq c n^{-\alpha/(2\alpha+1)}.$$

Idea of proof. The proof is based on Assouad's Lemma, as described in Tsybakov (2009) and the model to construct the lower bounds are closed to the one that are constructed in univariate non parametric regression. The main difficulty here is to handle the randomness of the empirical norm d_n itself (to the best of my knowledge lower bounds in empirical norms have not been proven yet so far). The idea is to work conditionally on the X_i's that gives us the intermediate result

$$\inf_{(\widehat{\beta},\widehat{g})} \sup_{\beta^* \in \mathcal{E}_b(R), \|g^*\|_\alpha \leq L} \mathbb{E}[d_n((\widehat{\beta},\widehat{g}),(\beta^*,g^*))] \geq cn^{-\alpha/(2\alpha+1)} \inf_{\widehat{\boldsymbol{\omega}} \in \Omega_n} \max_{\boldsymbol{\omega} \in \Omega_n} \mathbb{P}_{\boldsymbol{\omega}}(\{\widehat{\boldsymbol{\omega}} \neq \boldsymbol{\omega}\} \cap \mathcal{A}_n)$$

with Ω_n a subset of $\{0,1\}^m$ of cardinality larger than $2^{m/8}$ and such that the Hamming distance $\sum_{j=1}^m \mathbf{1}_{\omega_j \neq \omega'_j}$ between two elements $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)$ and $\boldsymbol{\omega}' = (\omega'_1, \ldots, \omega'_m)$ of Ω_n is larger than $m/8$ (the existence of such a set Ω_n is given by the Varshamov-Gilbert bound) and $\mathcal{A}_n$ a set, measurable w.r.t. $X_1, \ldots, X_n$ such that

$$\mathbb{P}(\mathcal{A}_n^c) \geq 1/2.$$

□

To obtain an optimal minimax result we need now to obtain the following lower-bound, which is still in progress,

$$\inf_{(\widehat{\beta},\widehat{g})} \sup_{\beta^* \in \mathcal{E}_b(R), \|g^*\|_\alpha \leq L} \mathbb{E}[d_n((\widehat{\beta},\widehat{g}),(\beta^*,g^*))] \geq cn^{-\min\{\alpha;1\}(b+\gamma)/(2(b+\gamma)+1)}.$$

The difficulty is to handle the term $\min\{\alpha;1\}$ appearing in the rate. A partial answer to this problem has been found with help of a recent article of Wibowo et al. (2020) that proves the existence of bi-Lipshitz continuous functions of order $\min\{\alpha;1\}$ in the case $\alpha < 1$ (the case $\alpha \geq 1$ has the simple example $t \mapsto (t+2)^2 - 4$) i.e. functions such that

$$c^{-1}|t-u|^{\min\{\alpha;1\}} \leq |\psi(t) - \psi(u)| \leq c|t-u|^{\min\{\alpha;1\}}, \qquad t,u \in [-1,1].$$

The randomness of the semi-norm can be treated with similar conditioning arguments than the previous lower bound.

Another difficulty due to the complexity of the definition of the prior distribution will come to obtain information on the posterior distribution in practice (on simulated or real data).

Obtaining frequentist estimators that achieve these rates should also be possible by defining projection estimators for g^* with bases adapted for the estimation on non-compact support such as the Hermite basis (Belomestny et al., 2019). Moreover, we also have to precise the exact identifiability condition in this model.

More general models than the single-models that should achieve non-parametric rates are the

multiple-index models:

$$Y = g^*(\langle \beta_1^*, X\rangle, \dots, \langle \beta_m^*, X\rangle) + \varepsilon,$$

where the multivariate link function $g^* : \mathbb{R}^d \to \mathbb{R}$ and the m indexes $\beta_1^*, \dots, \beta_m^*$ have to be estimated. In addition, the problem of selecting the number of indexes m also is open.

2.3.2 Achieving minimax risk in sparse multivariate FLR ?

The question of obtaining minimax rates in the multivariate functional linear regression model with a sparse slope vector of functions $\boldsymbol{\beta}^*$ is still open. First a lower bound has to be proven, maybe under a different assumption on the covariance operator $\boldsymbol{\Gamma}$ than the one of [RLasso].

Then a second step is to find an estimator that achieves this minimax rate. The question of optimality could be studied from a Bayesian point of view. Indeed, let us denote $|\cdot|_2$ (resp. $|\cdot|_1$) the 2 (resp. 1) norm of $\mathbb{R}^n$, the LASSO estimator

$$\widehat{\beta} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}|_2^2 + \lambda|\boldsymbol{\beta}|_1\}$$

in the classic multivariate linear model

$$\mathbf{Y} = X\beta^* + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I_n),$$

can be considered as the maximum of the posterior distribution (posterior mode) when the prior distribution Π on $\boldsymbol{\beta}^*$ is a product of Laplace densities of scale parameter λ. Castillo et al. (2015) proved in this context that the LASSO posterior distribution put no mass into balls of radius substantially larger than the minimax rates. They have also shown that the minimax posterior contraction rates can be achieved with an alternative prior distribution. In addition, the bias problem of the LASSO estimator – or, in other words, the fact that LASSO "underestimates" the large coefficients – mentioned in the simulation study in subsection 2.1.3 is widely discussed in the Bayesian literature. To solve this problem, several alternatives have been developed, such as spike-and-slab priors (see for example the review in Bai et al. 2021). A closer look at the literature on Bayesian model selection could lead to the development of an estimator that achieves a minimax optimal posterior concentration rate.

2.3.3 Taking account discretization and noise in regression models

Another perspective is to find the minimax rates in regression models when the functional data are not entirely observed. Indeed, assume as in [BPRR], that instead of observing directly $X(t)$ for all $t \in [0, 1]$, we observe $\{Z(t_j), i = 1, \dots, n; j = 1, \dots, p\}$ such that

$$Z(t_j) = X(t_j) + \eta_j,$$

where $(\eta_j)_{j=1,\dots,p}$ is an i.i.d. noise, independent of X_i and $t_1, \dots, t_p$ a fixed regular grid. Then, the estimation of β^* in the functional linear model

$$Y_i = \langle \beta^*, X_i \rangle + \varepsilon_i,$$

has to take into account two sources of noise and the uncertainty linked to the fact that we do not observe X outside the points of the grid. As in [BPRR], both lower and upper bounds on the risk then must depend on a double asymptotic (in the number of individuals n and in the number of points p). Smoothing splines estimators have been studied in this context by Crambes et al. (2009). They proved that taking into account the noise and discretization in the study of the risk of the estimator of β^* adds an extra variance term of order $(\rho np)^{-1}$ where ρ is the parameter appearing in the penality of their ridge estimation procedure. It could be interesting to study minimax optimal projection estimators in this context by combining the results of [BMR16] and [BPRR].

Chapter 3

Constructing adaptive estimators: case of functional and/or dependent data

In Chapter 2, we give minimax rates for regression models involving functional covariates. In all these examples, we are able to define either projection estimators $\widehat{\beta}_{m^*}$, $\beta^*_{m_l^*,m_r^*}$ or kernel estimators $\widehat{F}_{h^*}$, $\widehat{m}_{h^*}$ that attain the lower bound for the minimax risk. However, since the optimal values m^*, m_l^*, h^* depend on the regularity of the functional data X and/or the regularity of the function to estimate which are both unknown, it is of interest to develop data-driven selection procedures for these parameters such that the rate of the selected parameter is comparable, in some sense, to the optimal rate.

3.1 Model selection

3.1.1 Model selection and functional PCA

In [CR15] and [BMR16], we have adapted model selection procedures to select the dimension m of the estimator $\widehat{\beta}_m$ in the functional linear model $Y = \langle \beta^*, X\rangle_{\mathcal{X}} + \varepsilon$. The approach is inspired by Barron et al. (1999); Baraud (2000, 2002) and consists in penalizing the least-squares criterion (least square contrast),

$$\widehat{m} \in \arg\min_{m=1,\dots,N_n} \left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle \widehat{\beta}_m, X_i\rangle_{\mathcal{X}})^2 + \kappa \widehat{\sigma}^2_m \frac{m}{n} \right\} \tag{3.1}$$

where

$$\widehat{\sigma}^2_m = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle \widehat{\beta}_m, X_i\rangle_{\mathcal{X}})^2$$

is an estimator of the noise variance; or equivalently

$$\widehat{m} \in \arg\min_{m=1,\ldots,N_n} \left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle \widehat{\beta}_m, X_i\rangle_{\mathcal{X}})^2 \left(1+\kappa\frac{m}{n}\right)\right\} \tag{3.2}$$

where $\kappa \geq 4$ is a constant and N_n is the maximal dimension.

The motivation behind the introduction of the estimator $\widehat{\sigma}^2_m$ for the noise variance is its simplicity since the criterion takes a multiplicative form. The simulation results of Table 1 of [BMR16] indicates that the substitution of σ^2 by $\widehat{\sigma}^2_m$ leads to similar model selection and estimation performances. The estimation performances, in terms of predictive risk, also are similar to the one of leave-one-out cross-validation, with a computation time drastically reduced: for the calculation of (3.2) we have to calculate $\widehat{\beta}_m$ for all m whereas leave-one-out cross-validated criterion necessitates the calculation of a least-squares estimator for each element of the sample and for all m. From a theoretical viewpoint, we obtained the following oracle-type inequality.

Theorem 11 ([BMR16]). Under some assumptions on the scores $\langle X, \psi_j\rangle_{\mathcal{X}}$ and a joint constraint on the regularities of β^* and X (see Theorem 3 of [BMR16]),

$$\mathbb{E}[\|\widehat{\beta}_{\widehat{m}} - \beta^*\|^2_\Gamma] \leq C \min_{m=1,\ldots,N_n} \left(\|\beta^* - \Pi_{S_m^{PCA}}\beta^*\|^2_\Gamma + \sigma^2\frac{m}{n}\right) + \frac{C'}{n}(1+\|\beta^*\|^2). \tag{3.3}$$

Idea of proof. The proof is inspired from the proofs of Baraud (2000, 2002). The randomness of the projection space is handled by working conditionally to the covariate. This allows us to obtain an oracle-type inequality controlling the empirical semi-norm $\|\cdot\|_n = \|\widehat{\Gamma}^{1/2}\cdot\|$ instead of the prediction norm $\|\cdot\|_\Gamma = \|\Gamma^{1/2}\cdot\|$. The main difficulty is the replacement of the empirical semi-norm by the targeted prediction norm which is made by controlling uniformly the ratios

$$\inf_{f\in\widehat{S}_{N_n}} \frac{\|f\|^2_n}{\|f\|^2_\Gamma}.$$

It is usually done by controlling the spectral radius of Gram matrices. The randomness of the model space is handled with perturbation theory tools described in Section 1.1.2. □

This allows us to prove that $\widehat{\beta}_{\widehat{m}}$ achieves the minimax rates detailed in Table 2.2 with a completely data-driven dimension selection procedure. Then $\widehat{\beta}_{\widehat{m}}$ is an adaptive estimator.

Using the fact that the multivariate functional linear regression model is a functional linear model on the product space $\mathcal{X} = \prod_{j=1}^d \mathcal{X}_j$ a similar criterion has been used in [RLasso] in the context of multivariate functional linear regression

$$Y = \langle \beta_1^*, X^{(1)}\rangle_{\mathcal{X}_1} + \ldots + \langle \beta_d^*, X^{(p)}\rangle_{\mathcal{X}_p} + \varepsilon$$

to select the dimension m of the estimator $\widehat{\beta}_{\boldsymbol{\lambda},m}$ defined by Eq. (2.5), p. 27. The criterion

considered is

$$\widehat{m} \in \arg\min_{m=1,\dots,N_n} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{d}\langle \widehat{\beta}_{\boldsymbol{\lambda},m}, X^{(j)}\rangle_{\mathcal{X}_j}\right)^2 + \kappa\sigma^2\log(n)\frac{m}{n}\right\}. \tag{3.4}$$

Compared to (3.1), the penalty term has an extra log term which was necessary to obtain theoretical results.

Theorem 12. [RLasso] Assume $\boldsymbol{\lambda}$ is chosen as in Proposition 7, there exist a universal constant $C_{MS} > 0$ and a minimal value $\kappa_{\min}$ such that, with probability larger than $1 - d^{1-q} - C_{MS}/n$, for all $\kappa \geq \kappa_{\min}$, for all $\zeta > 0$,

$$\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}} - \boldsymbol{\beta}^*\right\|_n^2 \leq (1+\zeta)\min_{m=1,\dots,N_n}\min_{\boldsymbol{\beta}\in\mathcal{X}^{(m)},|J(\boldsymbol{\beta})|\leq s}\left\{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_n^2 + \frac{9}{4(\tilde{\kappa}_n^{(m)})^2}\sum_{j\in J(\boldsymbol{\beta})}\lambda_j^2 + \frac{2+\zeta}{1+\zeta}\kappa\sigma^2\log(n)\frac{m}{n}\right\}.$$

Compared to Proposition 7, there is an additional term due to the model selection penalty, which is negligible with respect to the term due to the sparsity inducing penalty. The proof is based on the control of the empirical process $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\sum_{j=1}^{d}\langle f_j, X^{(j)}\rangle_{\mathcal{X}_j}$ uniformly over the functions $\mathbf{f} = (f_1,\dots,f_d) \in \mathcal{X}^{(N_n)}$ and is also inspired from the proofs of Baraud (2000).

The resulting estimator is not adaptive, it achieves the same rate as the estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m^*}$ (see Eq. 2.9, p. 30) but this rate is (probably) not the minimax rate as explained in subsection 2.1.3. However, the criterion (3.4) gives us a fully data-driven method (after replacement of σ^2 by $\widehat{\sigma}_m^2$) which is also computationally efficient to select the dimension m with similar support selection performances. Hence, in the context of variable selection in multivariate FLR, it provides an advantageous alternative to cross-validation that is too time consuming to be efficiently used in practice in our context.

In the case of the functional linear model with functional output

$$Y = B^*X + \varepsilon,$$

we adopted in [CMR], a similar approach by selecting the two unknown dimensions $\widehat{m}_l$ and $\widehat{m}_r$ with the following criterion

$$(\widehat{m}_l, \widehat{m}_r) \in \arg\min_{m_l=1,\dots,N_n; m_r\in\mathbb{N}^*\cup\{+\infty\}} \frac{1}{n}\sum_{i=1}^{n}\|Y_i - BX_i\|^2 + \kappa\sigma^2\frac{m_l}{n},$$

and since the first term is decreasing with m_r and the penalty is independent of m_r we can choose $\widehat{m}_r = +\infty$. We also obtain an oracle-type inequality in empirical norm.

Theorem 13. [CMR] Under some assumptions on the eigenvalues sequence $(\lambda_j)_{j\geq 1}$, on the scores

$\xi_j = (\lambda_j^{-1/2}\langle X, \Phi_j\rangle)_{j\geq 1}$ and if $\|\varepsilon\|$ has a moment of order strictly larger than 6, then for all $\zeta > 0$,

$$\mathbb{E}\left[\left\|B^* - \widehat{B}_{\widehat{m}_l,+\infty}\right\|_n^2\right] \leq (1+\zeta) \inf_{m_l\in 1,\dots,N_n} \left\{\mathbb{E}[\|B^* - \Pi_{S_{m_l}^{PCA}\otimes\mathbb{L}^2([0,1])}B^*\|_n^2] + \frac{2+\zeta}{1+\zeta}\kappa\sigma^2\frac{m_l}{n}\right\} + \frac{C}{n}$$

where the empirical semi-norm writes $\|T\|_n^2 = \frac{1}{n}\sum_{i=1}^n \|TX_i\|^2$.

3.1.2 Estimation of hazard rate and multiplicative censoring [CCR17]

We consider in [CCR17] hazard rate estimation under multiplicative censoring. More precisely, let X be a real non-negative random variable (supposed to represent e.g. a duration), the hazard rate is the quantity

$$w^*(t) = \lim_{h\to 0} \mathbb{P}(X > t+h | X > t) = \frac{f_X(t)}{\bar{F}_X(t)}$$

of a real positive random variable X of density f_X and survival function $\bar{F}_X(t) = \mathbb{P}(X > t)$. If X represents a lifetime for instance, the hazard rate models the probability of dying "just after" time t conditionally to the fact that the person is alive at time t. Since the hazard rate is a quotient with the survival function that decreases to 0 when $t \to +\infty$, the behavior of its estimators may be unstable when t is large.

In [CCR17], we consider adaptive estimation in presence of multiplicative censoring that is to say we assume that we observe $Y_1, \dots, Y_n \sim_{i.i.d.} Y$ where Y is a random variable such that

$$Y = UX, \qquad U \sim \mathcal{U}([0,1]), \quad U \perp\!\!\!\perp X.$$

We estimate the target function w^* on a compact set $[0, \mathbf{a}]$, $\mathbf{a} > 0$, and define a collection of models $S_m = \text{span}\{\varphi_j, j \in \mathbb{J}_m\}$ where $\mathbb{J}_m$ is a subset of $\mathbb{Z}$. For instance, in the case of the B-splines model of order $r \in \mathbb{N}^*$, $\mathbb{J}_m = \{-r+1, \dots, 2^m - 1\}$, and

$$\varphi_{j,m}(t) = \frac{2^{m/2}}{\sqrt{\mathbf{a}}} N_r\left(\frac{2^m}{\mathbf{a}}t - j\right),$$

with $N_r(t) = \mathbf{1}_{[0,1]}^{\star r}(t)$ the indicator function on $[0,1]$ convoled r-times (i.e. N_r is the density function of the sum of r independent uniform variables on $[0,1]$). Drawing inspiration from Comte et al. (2011); Plancade (2011), we consider a minimum contrast estimator

$$\widehat{w}_m = \arg\min_{w\in S_m} \gamma_n(w) \quad \text{with} \quad \gamma_n(w) = \|w\|_n^2 - 2\nu_n(w),$$

where $\|\cdot\|_n$ (resp. ν_n) is an empirical semi-norm (resp. an empirical linear functional) adapted to our problem. The model is then chosen classically with a penalized contrast estimator

$$\widehat{m} \in \arg\min_{m=1,\dots,N_n} \{\gamma_n(\widehat{w}_m) + \kappa\widehat{\text{pen}}(m)\},$$

with $\widehat{\text{pen}}(m) = \frac{1}{\mathfrak{d}_1 \widehat{\bar{F}_Y(\mathbf{a})} n} \left(\phi_0^2 |S_m| + (\mathbf{a}\phi_0^2 + \mathbf{a}^2\phi_1^2)|S_m|^3\right)$ where $\mathfrak{d}_1$, ϕ_0 and ϕ_1 are explicit quantities depending on the basis and $\widehat{\bar{F}_Y(\mathbf{a})}$ is a consistent estimator of the survival function of Y at point $\mathbf{a}$ such that $\widehat{\bar{F}_Y(\mathbf{a})} \geq n^{-1/2}$. The term $\widehat{\text{pen}}(m)$ is an estimation of the variance term and we remark that it increases when $\mathbf{a}$ increases. In particular, if Y (hence X) is compactly supported on $[0, \mathbf{b}]$ we expect estimation to be unstable when $\mathbf{a}$ is close to $\mathbf{b}$.

This can be seen in the upper-bound (oracle-type inequality) we obtain on the selected estimator.

Theorem 14. [CCR17] We assume the density of X is upper-bounded on $[0, \mathbf{a}]$, the survival function $\bar{F}_X$ of X verifies $\bar{F}_X(\mathbf{a}) > 0$, that $\|f_Y\|_{\mathbb{L}^2([0,\mathbf{a}])} < +\infty$ and $\mathbb{E}[Y_1^2] < +\infty$. Then there exists κ_0 and n^* such that, for $\kappa > \kappa_0$, $n \geq n^*$,

$$\mathbb{E}[\|\widehat{w}_{\widehat{m}} - w^*\|_{\mathbb{L}^2([0,\mathbf{a}])}] \lesssim \min_{m=1,\dots,N_n; |S_m| \geq \ln(n)} \left\{ \inf_{w \in S_m} \|w - w^*\| + \text{pen}(m) \right.$$
$$\left. + \mathbb{E}\left[\max\{|S_{\widehat{m}}|; |S_m|\} \left(\widehat{f}_Y(\mathbf{a}) - f_Y(\mathbf{a})\right)^2 \right] \right\} + n^{-1}.$$

In the upper-bound, in addition to the usual bias-variance term, appears an additional term depending on the difference $\left(\widehat{f}_Y(\mathbf{a}) - f_Y(\mathbf{a})\right)^2$. Using a locally adaptive estimator for $f_Y(\mathbf{a})$, such as the one defined by Rebelles (2015), allows us to obtain a quantity that does not degrades the convergence rates.

3.2 Bandwidth selection for kernel estimators

The second large class of estimators for which we are able to define data-driven selection procedure leading to adaptive estimators are kernel estimators. Important recent progresses have been made from the seminal work of Goldenshluger and Lepski (2011). The general idea is the following: let

$$\widehat{\nu}_h(t) = \frac{1}{nh^d} \sum_{i=1}^{n} K_h(t - T_i),$$

a kernel estimator for the d-variate density ν of a sample $T_1, \dots, T_n \sim_{i.i.d} \nu$. The bandwidth is selected with the following criterion:

$$\widehat{h} \in \arg\inf_{h \in \mathcal{H}} \left\{ \sup_{\eta \in \mathcal{H}} \left(\|K_h \star \widehat{\nu}_\eta - \widehat{\nu}_h\| - m(h, \eta) \right)_+ + m^*(h) \right\}, \tag{3.5}$$

where the quantity $m(h, \eta)$, called majorant, is a term allowing to control uniformly in η and h the stochastic errors $\frac{1}{n}\sum_{i=1}^n K_h(t - T_i) - \mathbb{E}[K_h(t - T_1)]$ and $\frac{1}{n}\sum_{i=1}^n K_h \star K_\eta(t - T_i) - \mathbb{E}[K_h \star K_\eta(t - T_1)]$ and $m^*(h) = \sup_{\eta \in \mathcal{H}} m(\eta, h)$. The first term $\sup_{\eta \in \mathcal{H}} \left(\|K_h \star \nu_h - \nu_h\| - m(h, \eta) \right)_+$ is constructed to be of the order of the bias term $\|K_h \star f - f\|$ and replaces the term depending on the least-squares contrast for the model selection estimators of Section 3.1.1. The bandwidth collection

$\mathcal{H}$ is, in the article of Goldenshluger and Lepski (2011), a compact subset of $]0,+\infty[^d$.

Several simplifications, variations and adaptations to different contexts have been made to this initial work. A common simplification can be made by considering that the bandwidth collection $\mathcal{H}$ is finite with cardinality growing with n. This allows to control the stochastic error uniformly in n with the simple argument

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}K_h(t-T_i)-\mathbb{E}[K_h(t-T_1)]-m(h)\right)_+\right]$$
$$\leq\sum_{h\in\mathcal{H}}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}K_h(t-T_i)-\mathbb{E}[K_h(t-T_1)]-m(h)\right)_+\right].$$

Hence, the control of this term is based on an appropriate concentration inequality. Another simplification that can be found in numerous works on the subject is to choose the majorant $m(h,\eta)$ of the order of the variance term of $\widehat{\nu}_h$ (here $V(h)=\|K\|_2^2/(nh_1\times\ldots\times h_d)$).

3.2.1 Bandwidth selection for the invariant measure of a Bifurcative Markov Chain [BR20]

Starting from these considerations we consider adaptations of the Goldenshluger and Lepski's method in the case of Bifurcating Markov Chains. Bifurcating Markov Chains are a class of stochastic processes indexed by a binary tree $\mathbb{T}=\bigcup_{k=1}^{+\infty}\{0,1\}^k$ satisfying a Markov property. The precise definition is given below.

Definition 15 (Bifurcating Markov Chain). Let μ be a probability measure on $(\mathbb{R}^d,\mathcal{B}(\mathbb{R}^d))$, $\mathcal{P}$ a $\mathbb{T}$-transition probability (i.e. $x\mapsto\mathcal{P}(x,A)$ is a measurable map, for all $A\in\mathcal{B}(\mathbb{R}^d)\otimes\mathcal{B}(\mathbb{R}^d)$ and $A\mapsto\mathcal{P}(x,A)$ is a probability measure for all $x\in\mathbb{R}^d$) and $(\mathcal{F}_n)_{n\geq1}$ be a filtration. The process $(X_u)_{u\in\mathbb{T}}$ is called a $(\mathcal{F}_n)-BMC$ if

- X_u is $\mathcal{F}_n$-measurable for all $u\in\mathbb{G}_n=\{0,1\}^n$ (the n-th generation of the tree),
- $X_\emptyset\sim\mu$ (initial distribution).
- For all $n\in\mathbb{N}^*$, for all $(f_u)_{u\in\mathbb{G}_m}$ measurable functions from $(\mathbb{R}^d)^3$ to $\mathbb{R}$,

$$\mathbb{E}\left[\prod_{u\in\mathbb{G}_n}f_u(X_u,X_{u0},X_{u1})|\mathcal{F}_n\right]=\prod_{u\in\mathbb{G}_n}\int_{\mathbb{R}^d\times\mathbb{R}^d}f_u(X_u,y,z)\mathcal{P}(x,dy,dz),$$

  with $X_{u0}=(u,0)$ and $X_{u1}=(u,1)$ the values of the process on the two parents $(u,0)\in\mathbb{G}_{n+1}$ and $(u,1)\in\mathbb{G}_{n+1}$ of u.

These processes are of interest to sudy the propagation of physical caracteristics (size, weight,...) of individuals from a lineage.

From a BMC $(X_u)_{u\in\mathbb{T}}$, we can construct a Markov chain $(Y_n)_{n\geq 1}$ called tagged-branch chain started from $Y_0 = X_\emptyset$ and constructed iteratively by selecting independently one of the two parents with equal probability (if $Y_n = X_u$ with $u \in \mathbb{G}_n$, then $Y_{n+1} = X_{u0}$ with probability $1/2$ or $Y_{n+1} = X_{u1}$ with probability $1/2$ independently of $(X_u)_{u\in\mathbb{T}}$). Assuming the Markov chain $(Y_n)_{n\geq 1}$ is ergodic, we estimate its invariant distribution ν from the observation $(X_u)_{u\in\mathbb{T}_n}$ of the BMC until the n-th generation with the kernel estimator

$$\widehat{\nu}_h(x) = \frac{1}{|\mathbb{T}_n|}\sum_{u\in\mathbb{T}_n} K_h(x - X_u).$$

A bias-variance decomposition for the pointwise risk

$$\mathbb{E}[(\widehat{\nu}_h(x) - \nu(x))^2] \leq 2(K_h \star \nu(x) - \nu(x))^2 + 2\frac{C(P,\nu)}{|\mathbb{T}_n|h_1 \times h_d}$$

was obtained by Bitseki-Penda and Olivier (2017), with $C(\mathcal{P},\nu)$ a quantity depending on the distribution μ, on the transition probability $\mathcal{P}$ and on the kernel K.

Assuming $C(\mathcal{P},\nu)$ is known we defined in [BR20] a local bandwidth selection criterion of the form

$$\widehat{h}(x) \in \arg\min_{h\in\mathcal{H}} \left\{ \max_{\eta\in\mathcal{H}} \left((K_h \star \widehat{\nu}_\eta(x) - \widehat{\nu}_h(x))^2 - bV(h,x) \right)_+ + V(h,x) \right\},$$

with

$$V(h,x) = C(\mathcal{P},\mu)\frac{\log(|\mathbb{T}_n|)}{|\mathbb{T}_n|\ h_1 \dots h_d}$$

for which we prove an oracle-type inequality.

Theorem 16. Under an assumption of geometric uniform ergodicity of the BMC $(X_u)_{u\in\mathbb{T}_n}$, and if $\min_{h\in\mathcal{H}_n} h_1 \times \ldots \times h_d \geq \log(|\mathbb{T}_n|)/|\mathbb{T}_n|$,

$$\mathbb{E}[(\widehat{\nu}_{\widehat{h}(x)}(x) - \nu(x))^2] \leq C_1 \min_{h\in\mathcal{H}_n} \{\mathcal{B}_h(x) + V(x,h)\} + \frac{C_2}{|\mathbb{T}_n|},$$

with

$$\mathcal{B}_h(x) = \max_{\eta\in\mathcal{H}_n} ((K_h \star K_\eta \star \nu(x) - K_\eta \star \nu(x))^2.$$

The key arguments of the proof are Bernstein-type inequalities that we prove to control the two terms $\frac{1}{|\mathbb{T}_n|}\sum_{u\in\mathbb{T}_n} K_h(x - X_u) - \mathbb{E}[K_h(x - X_u)]$ and $\frac{1}{|\mathbb{T}_n|}\sum_{u\in\mathbb{T}_n} K_h \star K_\eta(x - X_u) - \mathbb{E}[K_h \star K_\eta(x - X_u)]$.

The term $\mathcal{B}_h(x)$ is of the order of the bias term for appropriate choices of $\mathcal{H}_n$. The form of the bias term $\mathcal{B}_h(x)$ is usual in non-parametric estimation. However, a similar form of the bias in oracle-type inequalities for the pointwise risk also appears in the case of density estimation for i.i.d. data (Rebelles, 2015) or in deconvolution problems (Comte and Lacour, 2013). An

immediate consequence is that, if ν is a $\beta = (\beta_1, \ldots, \beta_d)$-Hölder continuous density, then

$$\mathbb{E}[(\widehat{\nu}_{\widehat{h}(x)}(x) - \nu(x))^2] \leq C \left(\frac{\log(|\mathbb{T}_n|)}{|\mathbb{T}_n|} \right)^{-2\bar{\beta}/(2\bar{\beta}+1)},$$

where $\bar{\beta} = d/(1/\beta_1 + \ldots + 1/\beta_d)$ is the harmonic mean of β. This rate coincides with the minimax rate for density estimation for i.i.d. data under anisotropic regularity and then our estimator is adaptive.

Now from a practical viewpoint, the constants a and b appearing in the criterion fulfill the constraint $b \geq a \geq 1$ (and we choose, as suggested in Lacour and Massart 2016 $b = 2a$). The main difficulty is the fact that the quantity $C(\mathcal{P}, \mu)$ is not explicit raises a problem in practice. For the case $d = 1$, we adapt the method, developed by Arlot and Massart (2009) for least-squares estimators that detect complexity jumps to find a good value of $aC(\mathcal{P}, \mu)$. The whole procedure works well in practice (in dimension $d = 1$).

3.2.2 Bandwidth selection and functional data [CR14], [CR16]

Another variation of the criterion defined by Goldenshluger and Lepski (2014) in the case where the bias does not write as a convolution product consists in replacing the term $K_h \star \widehat{\nu}_\eta$ in (3.5) by $\widehat{\nu}_{\min\{\eta;h\}}$. This criterion has been considered in different context (for instance Rebelles 2015) and is, in fact, a rediscovery of Kerkyacharian et al. (2001). We consider a criterion of this type to select the bandwidth h for both estimators $\widehat{F}^x_{h,\mathfrak{d}}$ and $\widehat{m}^x_{h,\mathfrak{d}}$ of the conditional c.d.f and regression function respectively. The selected estimators achieve an oracle-type inequality and attain the minimax rates of Table 2.5, up to a $\log(n)$ term in the case $H_{X,pol}$.

3.3 Perspectives

3.3.1 Functional autoregressive processes [MR]

In this project, we aim at estimating the transition operator in a functional autoregressive process (functional AR(1) process) defined by

$$X_{k+1} = \Phi^* X_k + \varepsilon_{k+1}, \qquad k \in \mathbb{Z},$$

based on the observations $X_1, \ldots, X_n$. Here $\Phi^* : \mathcal{X} \to \mathcal{X}$ is an unknown linear operator which is the parameter of the model that we intend to estimate. This model has been widely studied in the book of Bosq (2000) and is identifiable on the condition $|||\Phi^*||| < 1$ which is assumed hereafter.

Assumptions	$\lambda_j \asymp j^{-2\gamma}$		$\lambda_j \asymp e^{-\gamma j}$	
	$\|\Phi^*\psi_j\|^2 \asymp j^{-2r}$	$\|\Phi^*\psi_j\|^2 \asymp e^{-rj}$	$\|\Phi^*\psi_j\|^2 \asymp j^{-2r}$	$\|\Phi^*\psi_j\|^2 \asymp e^{-rj}$
$\mathcal{R}_n(\widehat{\Phi}_{\rho^*})$	$n^{-\frac{2\gamma+2r-1}{2\gamma+2r}}$	$n^{-\frac{4\gamma}{4\gamma+1}}$	$\sim \frac{\log(n)}{n}$	
ρ^*	$n^{-\frac{2\gamma}{2\gamma+2r+1}}$	$n^{-\frac{2\gamma}{4\gamma+1}}$	$\frac{\log^{2b}(n)}{n}$	$\frac{\log(n)}{n}$

Table 3.1: Minimax rates and theoretical optimal value of ρ in the $AR(1)$ model (conjecture).

We consider ridge estimators of Φ^*:

$$\widehat{\Phi}_\rho \in \arg\min_{\Phi\in\mathcal{S}_2}\left\{\frac{1}{n}\sum_{i=1}^{k-1}\|\Phi(X_{k+1})-\Phi(X_k)\|^2+\frac{\rho}{2}\|\Phi\|_{HS}^2\right\},$$

where $\mathcal{S}_2$ is the space of Hilbert-Schmidt operators equipped with its usual norm $\|\cdot\|_{HS}$. It can be shown that $\widehat{\Phi}_\rho$ can be written explicitly

$$\widehat{\Phi}_\rho = \widehat{D}(\widehat{\Gamma}+\rho I)^{-1},$$

where $\widehat{D} : f \mapsto \frac{1}{n}\sum_{i=1}^{n-1}\langle X_k, f\rangle X_{k+1}$ is the empirical cross-covariance operator.

The estimator $\widehat{\Phi}_\rho$ also has been studied by Caponera and Panaretos (2022) who obtain an upper-bound on the p-Schatten norms. We consider a predictive risk

$$\mathcal{R}_n(\widehat{\Phi}) = \mathbb{E}[\|(\widehat{\Phi}-\Phi^*)(X_{n+1})\|^2],$$

and study the dependency of the minimax rates of convergence with the decreasing rate of the eigenvalues $(\lambda_j)_{j\geq 1}$ of the covariance operator Γ and the regularity of the function to estimate as it has been done for the functional linear model (Tables 2.1 and 2.2). In our case, it is the decreasing rate of the quantity $\|\rho^*\psi_j\|^2$ the characterizes the regularity of ρ^* and conjecture the rates in Table 3.1.

Since the optimal value ρ^* depends on unknown regularities of X and Φ^*, the aim is to define an adaptive estimator by defining a data-driven selection criterion for ρ. The main obstacle we are facing is to find a sufficiently sharp concentration inequality in our case where the data are dependent and functional. Moreover, adaptive procedures for Ridge estimators are not common. Up to our knowledge, only two works are related to the subject in the case of multivariate i.i.d. data Loubes and Ludeña (2008); Baraud et al. (2014) and the path to follow here will certainly differ from these previous works.

3.3.2 Estimation of quantiles [CRS]

The aim is to define an adaptive estimator of the quantile of a distribution of a real random variable Y conditionally to a real random variable Z from a sample $\{(Y_i, Z_i), i = 1, \ldots, n\}$. We consider as in Guerre and Sabbah (2012) a local polynomial estimator: $\widehat{Q}_h(\alpha|z)$ is constructed

as the first element q_0 of the minimizer over $q = (q_0, \dots, q_\ell)^t \in \mathbb{R}^{\ell+1}$ of the criterion

$$\sum_{i=1}^{n} \ell_\alpha(Z_i - \mathbf{U}(X_i - x)^t q) K_h(X_i - x),$$

where K is a kernel function, $K_h(\cdot) = K(\cdot/h)/h$, $\ell_\alpha(z) = z(\alpha - \mathbf{1}_{z \le 0})$ and $\mathbf{U}(x) = (1, x, \dots, x^p/p!)^t$. The choice of the estimator is motivated by the fact that the quantile of order α of Y given Z is the solution of the minimization problem

$$\min_{q \in \mathbb{R}} \{\mathbb{E}[\ell_\alpha(Z - q) | X = x]\}.$$

The local polynomial estimator is necessary to achieve the right order of the bias term when the function to estimate is b-Hölder continuous, with $b > 1$.

Chichignoud and Loustau (2015) have developed an adaptive estimation procedure based on Goldenshluger and Lepski (2011) for estimators defined by minimization of an empirical risk but their methods requires that the criterion to minimize is 2-times differentiable, which is not the case here and adds some difficulties in obtaining the theoretical results.

3.3.3 Estimation of the ergodicity parameter in Markov Chains and BMC

The uniform ergodicity assumption we made in [BR20] to obtain the theoretical results is questionable. More precisely, we assume that there exists a constant $\rho \in (0, 1/2)$ and $M > 0$ such that

$$|\langle g, \mathcal{Q}^n \mu \rangle - \langle g, \nu \rangle| \le M \|f\|_\infty \rho^n, \qquad g : \mathbb{R}^d \to \mathbb{R}, \text{ bounded and } \nu - \text{integrable},$$

with $\mathcal{Q}^m \mu$ the distribution of Y_n.

The fact that $\rho < 1/2$ is itself a problem. In particular, there is no method to verify it on observations. Moreover, it seems that we have a phase-transition at $\rho = 1/2$: in the case $\rho \ge 1/2$, the rates of convergence are no the same than in the i.i.d. case (Bitseki-Penda and Delmas, 2022) and the optimal bandwidth may depend on ρ. Obtaining an adaptive estimator in this context hence requires to be able to estimate the parameter ρ. With Marc Hoffmann and Valère Bitseki-Penda, we have the project to develop an estimation procedure of the ergodicity parameter ρ, based on the consideration that ρ is the second eigenvalue of the integral operator $\mathcal{Q}$ associated to the transition kernel of the Markov chain $(Y_n)_{n \ge 1}$. Since this transition kernel is estimable from the observations, the approaches and tools described in Chapter 1, could be useful to study estimation of ρ.

3.3.4 Adaptive estimation of a "regular" density conditionally to a functional data [CR]

In [CR14] and [CR16], we estimate the conditional c.d.f. and regression function respectively with the assumption that the target functions are b-Hölder continuous, with $b \leq 1$. This strong assumption is necessary since we need the kernel to be positive. More precisely we need the existence of a constant $c \geq 1$ such that

$$c^{-1}\mathbf{1}_{[0,1]}(t) \leq K(t) \leq c\mathbf{1}_{[0,1]}(t), \qquad t \in \mathbb{R}.$$

To overcome this difficulty, we would like to use multiplicative kernels. Consider e.g. the estimation of the conditional density $f_{Y|X=x}$ w.r.t. the Lebesgue measure of a real random variable Y conditionally to $X = x$ where X is a functional variable. We can define an estimator of $f_{Y|X=x}$ from a sample $\{(X_i, Y_i), i = 1, \dots, n\}$ with the following formula

$$\widehat{f}_{h,w,m}(x) = \sum_{i=1}^{n} W_{h,p}^{(i)}(x) H_w(y - Y_i) \text{ with } W_{h,p}^{(i)}(x) = \frac{\prod_{j=1}^{m} K_{h_j}(\langle x - X_i, e_j \rangle)}{\sum_{i'=1}^{n} \prod_{j=1}^{m} K_{h_j}(\langle x - X_{i'}, e_j \rangle)},$$

where $h \in]0; +\infty[^m$, $w > 0$ and $m \in \mathbb{N}^*$ are the parameters to select and K and H are two kernels. An alternative would be to consider local polynomial estimators but writing such estimators in a functional context is not an easy task.

3.3.5 Bandwidth selection and EM algorithm [BCCHLR]

We consider here classification for multivariate data. Let X be a random variable in $\mathbb{R}^d$. We suppose that X follows the following mixing model i.e. that its density f_X can be written

$$f_X(t) = \sum_{k=1}^{K} p_k f_k(x), \tag{3.6}$$

where the weights $p_1, \dots, p_K \in [0, 1]$ verifies $\sum_{k=1}^{K} p_k = 1$ and the density of the classes $f_1, \dots, f_K$ are general multivariate densities in $\mathbb{R}^d$. The case that is of interest for us is the case $d = 2$ and $K = 3$ since the aim is to compare genetic expressions under two different conditions as in Bérard et al. (2011) without the assumption of gaussianity of the logarithm of the data.

Written in the general form (3.6), the model is not identifiable. Hence, we assume the existence of an unknown function $f : \mathbb{R} \to \mathbb{R}$, such that,

$$f_k(x_1, \dots, x_d) = \prod_{j=1}^{d} f(x_j - \mu_{j,k}),$$

where $(\mu_{j,k})_{j=1,\dots,d;k=1,\dots,K}$ are translation parameters that are also unknown. In other words, conditionally to the belonging to a class k, the coordinates $(X^1, \dots, X^d)$ of X are independent

and follow, up to a translation, the same distribution. A less restrictive model would be to also add a scaling parameter that may be different among the coordinates and/or the classes. However, adding this scaling parameter implies non identifiability of the model. On the contrary, Hunter et al. (2007) proves identifiability of the model with translation in the case of interest ($d = 2$, $m = 3$) under some mild conditions on the p_k's.

EM-like algorithms with kernel density estimators have been implemented in the R package Benaglia et al. (2009) with a bandwidth selection method based on Rule-of-Thumb, which is designed for Gaussian distribution but has no theoretical support for general distributions. Then, our idea was to replace the Rule-of-Thumb method by a bandwidth selection step based on the Penalized Comparizon to Overfitting (PCO) method developed by Lacour et al. (2017).

This leads us to Algorithm 3 and our preliminary results compared to the case where the bandwidth selection step (in blue) is replaced by the Rule-of-Thumb are given in Fig. 3.1.

Indeed, the bandwidth selection step of algorithm Benaglia et al. (2009) is based on Rule of Thumb method. Cross-validation methods are too time-consuming to be inserted in an EM-like algorithm. The advantage of the PCO method is that is easily implementable, advantageous in terms of computation times and covered by theoretical guarantees. The other novelty of our approach is that the bandwidth grid $\mathcal{H}$ varies along the iterations: $\mathcal{H}^{(0)} = \mathcal{H}$ where $\mathcal{H}$ is a large grid containing $h_{\min} = \|K\|_\infty (nKd)^{-1}$ and then $\mathcal{H}^{(\ell)}$ only contains the neighbors of $h^{(\ell-1)}$ in $\mathcal{H}$.

The results we obtained are represented in Figure 3.1.

The performances of Rule-of-Thumb and PCO-like methods are comparable except in the last case where our method perfoms best.

Algorithm 3 EM-like algorithm with data-driven bandwidth selection [BCCHLR]

Initialization: first classification made by hierarchical clustering to initialize $\widehat{z}_{i,k}^{(0)}$ ($\widehat{z}_{i,k}^{(\ell)}$ estimate at step ℓ $\mathbb{P}(X_i \text{ in class } k)$), $\widehat{\mu}_{j,k}^{(0)}$, $\widehat{f}^{(0)}$ and

$$\widehat{\lambda}_k^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{z}_{i,k}^{(0)}.$$

repeat

E-step: update the probability of i being in class k, for all $i = 1, \ldots, n$; $k = 1, \ldots, K$,

$$\widehat{z}_{i,k}^{(\ell)} = \frac{\widehat{p}_k^{(\ell-1)} \prod_{j=1}^{d} \widehat{f}^{(\ell-1)}(X_i - \widehat{\mu}_{j,k}^{(\ell-1)})}{\sum_{k'=1}^{m} \widehat{p}_{k'}^{(\ell-1)} \prod_{j=1}^{d} \widehat{f}^{(\ell-1)}(X_i - \widehat{\mu}_{j,k'}^{(\ell-1)})}$$

M like step:

1. Update the estimation of the weights: for all $k = 1, \ldots, K$

$$\widehat{p}_k^{(\ell)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{z}_{i,k}^{(\ell)}.$$

2. Update the estimation of translation parameters: for all $j = 1, \ldots, d$; $k = 1, \ldots, K$

$$\widehat{\mu}_{j,k}^{(\ell)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{z}_{i,k}^{(\ell)}}{\widehat{p}_k^{(\ell)}} X_i^j.$$

3. Update the estimation of the common density :

   (a) Calculate, for all $h \in \mathcal{H}^{(\ell-1)} \cup \{h_{\min}\}$,

$$\widehat{f}_h^{(\ell)}(t) = \frac{1}{nd}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{d} z_{i,k}^{(\ell)} K_h(t - X_i^k + \widehat{\mu}_{j,k}^{(\ell)}).$$

   (b) Select the bandwidth:

$$\widehat{h}^{(\ell)} = \arg\min_{h \in \mathcal{H}^{(\ell)}} \left\{ \|\widehat{f}_h - \widehat{f}_{h_{\min}}\|^2 + 2\frac{\langle K_h, K_{h_{\min}}\rangle}{n} \right\}.$$

until $\sum_{i=1}^{n}\sum_{k=1}^{K} |\widehat{z}_{i,k}^{(\ell)} - \widehat{z}_{i,k}^{(\ell-1)}| \leq s$ or maximal number of iterations reached

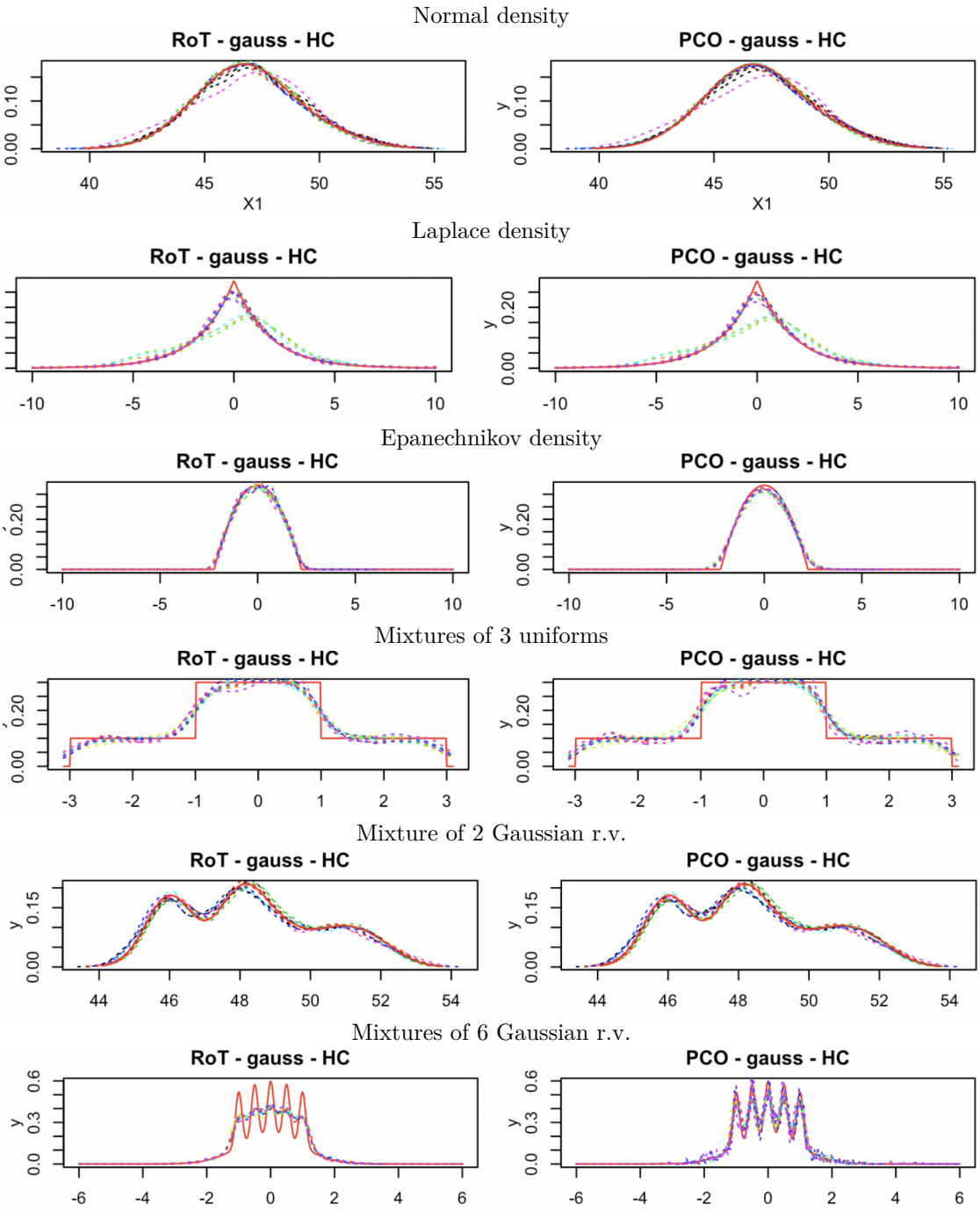

Figure 3.1: Comparison of EM-like algorithms in the case where the bandwidth is selected by the Rule-of-Thumb implemented in Benaglia et al. (2009)(left) or with the PCO method (inspired from Lacour et al. (2017)) with a Gaussian kernel $K(t) = (2\pi)^{-1/2}e^{-t^2/2}$. The true density is plotted in red.

Bibliography

S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. J. Mach. Learn. Res., 10:245–279, 2009.

J.-M. Azaïs and J.-C. Fort. Remark on the finite-dimensional character of certain results of functional statistics. C. R. Math. Acad. Sci. Paris, 351(3-4):139–141, 2013.

F. R. Bach. Consistency of the group lasso and multiple kernel learning. J. Mach. Learn. Res., 9:1179–1225, 2008.

R. Bai, V. Ročková, and E.I. George. Handbook of Bayesian Variable Selection, chapter Spike-and-slab meets LASSO: A review of the spike-and-slab LASSO, pages 81–108. Chapman Hall/CRC Press, 2021.

A. Bakhta, T. Boiveau, Y. Maday, and O. Mula. Epidemiological forecasting with model reduction of compartmental models. application to the covid-19 pandemic. Biology, 10(1):22, 2021.

Y. Baraud. Model selection for regression on a fixed design. Probab. Theory Relat. Fields, 117 (4):467–493, Aug 2000.

Y. Baraud. Model selection for regression on a random design. ESAIM Probab. Statist., 6: 127–146 (electronic), 2002.

Y. Baraud, C. Giraud, and S. Huet. Estimator selection in the Gaussian setting. Ann. Inst. Henri Poincaré Probab. Stat., 50(3):1092–1119, 2014.

A. Barron, L. Birgé, and P.l Massart. Risk bounds for model selection via penalization. Probab. Theory Relat. Fields, 113(3):301–413, Feb 1999.

T. Bazin, A. Krebs, A. Jobart-Malfait, V. Camilo, V. Michel, Y. Benezeth, F. Marzani, E. Touati, and D. Lamarque. Multimodal imaging as optical biopsy system for gastritis diagnosis in humans, and input of the mouse model. eBioMedicine, 69:103462, 2021.

R. Belhakem. Étude statistique de l'analyse en composantes principales fonctionnelle dans les cadres uni et multivarié. PhD thesis, Université Paris Sciences et Lettres, 2022.

P. Bellec and A. Tsybakov. Bounds on the prediction error of penalized least squares estimators with convex penalty. In Vladimir Panov, editor, Modern Problems of Stochastic Analysis and Statistics, pages 315–333, Cham, 2017. Springer International Publishing. ISBN 978-3-319-65313-6.

D. Belomestny, F. Comte, and V. Genon-Catalot. Sobolev-Hermite versus Sobolev nonparametric density estimation on $\mathbb{R}$. Ann. Inst. Statist. Math., 71(1):29–62, 2019.

T. Benaglia, D. Chauveau, D. R. Hunter, and D. Young. mixtools: An R package for analyzing finite mixture models. Journal of Statistical Software, 32(6):1–29, 2009. URL https://www.jstatsoft.org/v32/i06/.

C. Bérard, M.-L. Martin-Magniette, V. Brunaud, S. Aubourg, and S. Robin. Unsupervised classification for tiling arrays: Chip-chip and transcriptome. Stat. Appl. Genet. Mol. Biol., 10 (1), 2011.

S. V. Bitseki-Penda and J.-F. Delmas. Central limit theorem for bifurcating Markov chains under pointwise ergodic conditions. Ann. Appl. Probab., 32(5):3817 – 3849, 2022.

S. V. Bitseki-Penda and A. Olivier. Autoregressive functions estimation in nonlinear bifurcating autoregressive models. Stat. Inference Stoch. Process., 20(2):179–210, 2017.

D. Bosq. Linear Processes in Function Spaces: Theory and Applications. Springer New York, 2000.

J. C. Bronski. Small ball constants and tight eigenvalue asymptotics for fractional Brownian motions. J. Theoret. Probab., 16(1):87–100, 2003.

E Brunel, F. Comte, and C. Lacour. Minimax estimation of the conditional cumulative distribution function. Sankhya A, 72(2):293–330, 2010.

T. T. Cai and M. Yuan. Optimal estimation of the mean function based on discretely sampled functional data: phase transition. Ann. Statist., 39(5):2330–2355, 2011.

T. T. Cai and M. Yuan. Minimax and adaptive prediction for functional linear regression. J. Amer. Statist. Assoc., 107(499):1201–1216, 2012.

A. Caponera and V. M. Panaretos. On the rate of convergence for the autocorrelation operator in functional autoregression. Statistics & Probability Letters, 189:109575, 2022.

H. Cardot and J. Johannes. Thresholding projection estimators in functional linear models. J. Multivariate Anal., 101(2):395–408, 2010.

H. Cardot and P. Sarda. Functional linear regression. In The Oxford handbook of functional data analysis, pages 21–46. Oxford Univ. Press, Oxford, 2011.

Ricardo Carrizo Vergara. Karhunen-Loève expansion of random measures. arXiv:2203.14202, 2022.

I. Castillo, J. Schmidt-Hieber, and A. van der Vaart. Bayesian linear regression with sparse priors. The Annals of Statistics, 43(5):1986 – 2018, 2015.

M. Chichignoud and S. Loustau. Bandwidth selection in kernel empirical risk minimization via the gradient. Ann. Statist., 43(4):1617–1646, 2015.

F. Comte and J. Johannes. Adaptive estimation in circular functional linear models. Math. Methods Stat., 19(1):42–63, 2010.

F. Comte and J. Johannes. Adaptive functional linear regression. Ann. Statist., 40(6):2765–2797, 2012.

F. Comte and L. Lacour. Anisotropic adaptive kernel deconvolution. Ann. Inst. Henri Poincaré Probab. Stat., 49:569–609, 2013.

F. Comte, S. Gaïffas, and A. Guilloux. Adaptive estimation of the conditional intensity of marker-dependent counting processes. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, 47(4):1171 – 1196, 2011.

C. Crambes and A. Mas. Asymptotics of prediction in functional linear regression with functional outputs. Bernoulli, 19(5B):2627–2651, 2013.

C. Crambes, A. Kneip, and P. Sarda. Smoothing splines estimators for functional linear regression. Ann. Statist., 37(1):35 – 72, 2009.

D. J. Daley and D. Vere-Jones. An introduction to the theory of point processes. Vol. II. Probability and its Applications (New York). Springer, New York, second edition, 2008. General theory and structure.

L. Di Domenico, G. Pullano, C. E. Sabbatini, P.-Y. Boëlle, and V. Colizza. Expected impact of lockdown in île-de-france and possible exit strategies. medRxiv, 2020.

T. Dunker, M. A. Lifshits, and W. Linde. Small deviation probabilities of sums of independent random variables. In Ernst Eberlein, Marjorie Hahn, and Michel Talagrand, editors, High Dimensional Probability, pages 59–74, Basel, 1998. Birkhäuser Basel. ISBN 978-3-0348-8829-5.

F. Ferraty and P. Vieu. Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. C. R. Acad. Sci. Paris Sér. I Math., 330(2):139–142, 2000.

F. Ferraty and P. Vieu. Nonparametric functional data analysis. Springer Series in Statistics. Springer, New York, 2006. Theory and practice.

F. Ferraty, A. Laksaci, and P. Vieu. Estimating some characteristics of the conditional distribution in nonparametric functional models. Stat. Inference Stoch. Process., 9(1):47–76, 2006.

F. Ferraty, J. Park, and P. Vieu. Estimation of a functional single index model. In Frédéric Ferraty, editor, Recent Advances in Functional Data Analysis and Related Topics, pages 111–116, Heidelberg, 2011. Physica-Verlag HD. ISBN 978-3-7908-2736-1.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. J. Statist. Softw., 33(1):1–22, 2010.

C. Giraud. Introduction to high-dimensional statistics, volume 139 of Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, FL, 2015.

A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. Ann. Statist., 39(3):1608–1632, 2011.

A. Goldenshluger and O. Lepski. On adaptive minimax density estimation on $\mathbb{R}^d$. Probab. Theory Relat. Fields, 159:479–543, 2014.

Emmanuel Guerre and Camille Sabbah. Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function. Econometric Theory, 28(1): 87–129, 2012.

W. Härdle, V. Spokoiny, and S. Sperlich. Semiparametric single index versus fixed link function modelling. Ann. Statist., 25(1):212 – 243, 1997.

J. Hoffmann-Jørgensen, L. A. Shepp, and R. M. Dudley. On the lower tail of gaussian seminorms. Ann. Probab., 7(2):319–342, 1979. ISSN 00911798.

T. Hsing and R. Eubank. Theoretical foundations of functional data analysis, with an introduction to linear operators. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2015.

D. R. Hunter, S. Wang, and T. P. Hettmansperger. Inference for mixtures of symmetric distributions. Ann. Statist., 35(1):224–251, 2007.

M. Imaizumi and K. Kato. PCA-based estimation for functional linear regression with functional responses. J. Multiv. Anal., 163:15–36, 2018.

G.M. James, J. Wang, and J. Zhu. Functional linear regression that's interpretable. Ann. Statist., 37(5A):2083–2108, 2009.

E. Karasözen, E. Nissen, P. Büyükakpınar, M. D. Cambaz, M. Kahraman, E. Kalkan Ertan, B. Abgarmi, E. Bergman, A. Ghods, and A. A. Özacar. The 2017 July 20 Mw 6.6 Bodrum–Kos earthquake illuminates active faulting in the Gulf of Gökova, SW Turkey. Geophys. J. Int., 214(1):185–199, 03 2018.

G. Kerkyacharian, O. Lepski, and D. Picard. Nonlinear estimation in anisotropic multi-index denoising. Probab. Theory Related Fields, 121(2):137–170, 2001.

C. Lacour and P. Massart. Minimal penalty for the goldenshluger-lepski method. Stoch. Process. Their Appl., 126:3774–3789, 2016.

C. Lacour, P. Massart, and V. Rivoirard. Estimator selection: a new method with applications to kernel density estimation. Sankhya A, 79(2):298–335, 2017.

G. R. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A statistical framework for genomic data fusion. Bioinformatics, 20(16):2626–2635, 2004.

J.-M. Loubes and C. Ludeña. Adaptive complexity regularization for linear inverse problems. Electron. J. Stat., 2:661–677, 2008.

K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. Ann. Statist., 39(4):2164–2204, 2011.

A. Mas. Lower bound in regression for functional data by representation of small ball probabilities. Electron. J. Stat., 6(none):1745 – 1778, 2012.

T. Masak, S. Sarkar, and V M Panaretos. Separable expansions for covariance estimation via the partial inner product. Biometrika, 110(1):225–247, 06 2022. ISSN 1464-3510.

Z. Naulet and J. Rousseau. Posterior concentration rates for mixtures of normals in random design regression. Electron. J. Stat., 11(2):4065 – 4102, 2017.

S. Plancade. Model selection for hazard rate estimation in presence of censoring. Metrika, 74 (3):313–347, 2011.

J. Ramsay and B. W Silverman. Functional Data Analysis. Springer New York, 2010.

G. Rebelles. Pointwise adaptive estimation of a multivariate density under independence hypothesis. Bernoulli, 20(9):1984–2023, 2015.

W. Rudin. Real and complex analysis. McGraw-Hill Book Co., New York, 1966.

A. B. Tsybakov. Introduction to nonparametric estimation. Springer Series in Statistics. Springer, New York, 2009.

S Wibowo, V Y Kurniawan, and Siswanto. The relation between hölder continuous function of order $\alpha \in (0, 1)$ and function of bounded variation. Journal of Physics: Conference Series, 1490(1):012043, mar 2020.

Y. Yang and H. Zou. A fast unified algorithm for solving group-lasso penalize learning problems. Stat. Comput., 25(6):1129–1141, 2015.

A. Zettl. Sturm-Liouville theory, volume 121 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2005.

V. Zipunnikov, B. Caffo, D. M. Yousem, C. Davatzikos, B. S. Schwartz, and C. Crainiceanu. Multilevel functional principal component analysis for high-dimensional data. J. Comput. Graph. Stat., 20(4):852–873, 2011.

RÉSUMÉ

Ce mémoire d'HDR s'articule en trois parties :

- La réduction de dimension pour données fonctionnelles et son extension aux processus de comptage. Le cœur de cette partie consiste notamment en l'étude de propriétés théoriques de l'Analyse en Composantes Principales fonctionnelle sous différents schémas d'observation.
- Les vitesses minimax en régression pour données fonctionnelles. Plusieurs modèles de régression ayant pour covariable au moins une donnée fonctionnelle sont étudiés. Dans ces modèles, le risque minimax dépend en particulier, en plus de la régularité de la fonction à estimer, de la régularité des données. Cette régularité peut-être caractérisée théoriquement soit via le comportement asymptotique des valeurs propres de l'opérateur de covariance, soit via le comportement asymptotique des probabilités de petite boule.
- L'estimation adaptative pour données fonctionnelles et/ou dépendantes. Dans cette partie, nous étudions l'extension de méthodes de sélection de modèle pour des estimateurs par projection ou de fenêtre pour des estimateurs à noyau dans différents cadres faisant intervenir des données fonctionnelles et/ou dépendantes.

ABSTRACT

This habilitation thesis is structured into three parts:

- Dimension reduction for functional data and its extension to counting processes. The core of this section primarily involves the study of theoretical properties of Functional Principal Component Analysis under various observation schemes.
- Minimax rates in regression for functional data. Different regression models with at least one functional data covariate are studied. The minimax rate depends on the regularity of the function to be estimated as well as the regularity of the data. This regularity can be theoretically characterized either via assumptions on the asymptotic behavior of the eigenvalues of the covariance operator or on the asymptotic behavior of small ball probabilities.
- Adaptive estimation for functional and/or dependent data. In this section, we investigate the extension of model selection methods for projection estimators and bandwidth selection methods for kernel estimators in various frameworks involving functional and/or dependent data.