UNIVERSITÉ DE MONTPELLIER

INSTITUT MONPELLIÉRAIN ALEXANDER GROTHENDIECK

MÉMOIRE D'HABILITATION À DIRIGER DES RECHERCHES

Présenté par

Christophe Crambes

CONTRIBUTIONS AUX MODÈLES FONCTIONNELS DE RÉGRESSION

Soutenu le 30 juin 2020 devant le jury composé de

Delphine Blanke	-	Université d'Avignon	-	*Examinatrice*
Élodie Brunel-Piccinini	-	Université de Montpellier	-	*Examinatrice*
Hervé Cardot	-	Université de Dijon	-	*Examinateur*
Frédéric Ferraty	-	Université de Toulouse	-	*Rapporteur*
Aloïs Kneip	-	Université de Bonn	-	*Rapporteur*
André Mas	-	Université de Montpellier	-	*Examinateur*
Cristian Preda	-	Université de Lille	-	*Rapporteur*
Pascal Sarda	-	Université de Toulouse	-	*Examinateur*

UNIVERSITÉ DE MONTPELLIER

INSTITUT MONPELLIÉRAIN ALEXANDER GROTHENDIECK

MÉMOIRE D'HABILITATION À DIRIGER DES RECHERCHES

Présenté par

Christophe Crambes

CONTRIBUTIONS AUX MODÈLES FONCTIONNELS DE RÉGRESSION

Soutenu le 30 juin 2020 devant le jury composé de

Delphine Blanke	-	Université d'Avignon	-	*Examinatrice*
Élodie Brunel-Piccinini	-	Université de Montpellier	-	*Examinatrice*
Hervé Cardot	-	Université de Dijon	-	*Examinateur*
Frédéric Ferraty	-	Université de Toulouse	-	*Rapporteur*
Aloïs Kneip	-	Université de Bonn	-	*Rapporteur*
André Mas	-	Université de Montpellier	-	*Examinateur*
Cristian Preda	-	Université de Lille	-	*Rapporteur*
Pascal Sarda	-	Université de Toulouse	-	*Examinateur*

REMERCIEMENTS

Je voudrais tout d'abord remercier Frédéric Ferraty, Aloïs Kneip et Cristian Preda qui m'ont fait l'honneur d'accepter de rapporter ce manuscrit. Merci pour votre travail minutieux et vos commentaires enthousiastes. Je remercie également Delphine Blanke d'avoir gentiment accepté de faire partie du jury et de s'être rendue disponible.

J'adresse des remerciements tout particuliers à Pascal Sarda et Hervé Cardot pour leur participation à ce jury. Ce sont eux qui ont guidé mes tout premiers pas dans la recherche. Tout a commencé en DESS à Toulouse il y a presque 20 ans. Le temps passe à une vitesse incroyable, mais je n'oublie pas que, si j'en suis arrivé là aujourd'hui, vous y avez largement contribué.

Je remercie aussi André Mas et Élodie Brunel de faire partie du jury. Merci pour votre soutien qui m'a permis d'aborder la soutenance plus sereinement. Merci à Jean-Noël Bacro, qui m'a encouragé à réaliser ce travail de synthèse. L'Institut Montpelliérain Alexander Grothendieck et l'équipe de Probabilités et Statistiques offrent un cadre de travail stimulant et épanouissant, je remercie tous les collègues que j'ai la chance d'y cotoyer, ainsi que le personnel administratif et technique qui facilite grandement le quotidien.

Les travaux que je présente dans ce mémoire n'auraient pas existé sans les personnes avec lesquelles j'ai collaboré. Je remercie Aboubacar Amiri, Ali Gannoun, Yousri Henchiri, Nadine Hilgert, Tito Manrique, André Mas et Baba Thiam pour ces rencontres et ces échanges, tant sur le plan scientifique que sur le plan humain.

Je voudrais également remercier mes collègues de l'IUT. Être dans ce département d'enseignement m'a amené à rencontrer des collègues issus d'autres disciplines, je n'imaginais pas en arrivant en TC à quel point ce serait enrichissant et stimulant. Un grand merci en particulier à Sophie, Audrey et Michèle, c'est une chance et un plaisir de travailler avec vous.

Mes derniers remerciements iront à ma famille. Je remercie mes parents, ainsi que Magali et Marc d'avoir toujours été là pour moi. Enfin, j'ai la chance incroyable d'être un mari et un papa comblé. Merci à Marine, Théa et Zoé pour tout le bonheur que vous m'apportez. Je vous aime.

RÉSUMÉ

Mes recherches portent sur la statistique fonctionnelle. Cette branche de la statistique étudie des modèles lorsque les données sont assimilables à des courbes. En effet, avec les moyens technologiques actuels, certaines données sont relevées sur des grilles de mesure très fines. De plus, ces données proviennent fréquemment de mesures de phénomènes de nature continue (par exemple des relevés de températures, des courbes de croissance, ...), et il est naturel de les traiter en tant que fonctions (du temps, de l'espace) plutôt qu'en tant que vecteurs de points de mesures. On considère alors que l'on observe des réalisations de variables aléatoires à valeurs dans un certain espace de fonctions.

Mes recherches se sont essentiellement focalisées sur la régression mettant en jeu des variables fonctionnelles. Le point de départ est le modèle le plus simple, le modèle de régression linéaire fonctionnelle, dans lequel la variable d'intérêt est à valeurs réelles, et la variable explicative est à valeurs dans un espace de fonctions. Je me suis alors intéressé à différentes extensions de ce modèle.

La première extension introduit des données manquantes dans les observations. Dans un premier temps, je me suis intéressé à une méthode d'imputation de données manquantes sur la variable réponse, afin de reconstituer un jeu de données complet. Ce jeu de données complété permettra une prévision ultérieure lorsque qu'apparaît une nouvelle observation de la variable explicative. Des prolongements de ce travail sont envisagés, notamment le fait de considérer des données manquantes à la fois sur la variable explicative et la variable d'intérêt, ainsi qu'une variable d'intérêt elle aussi à valeurs dans un espace de fonctions.

La deuxième extension considère une variable réponse à valeurs dans un espace de fonctions. Ce travail introduit une méthode d'estimation du paramètre fonctionnel du modèle et établit notamment des résultats asymptotiques sur l'erreur de prévision sur la réponse. En parallèle, je me suis également intéressé à un sous-modèle du modèle précédent dans le cas (notamment lorsque les variables sont fonctions du temps) où la valeur de la variable réponse en un certain instant est expliquée par le passé de la variable explicative avant cet instant.

La troisième extension consiste à se placer dans un cadre non-paramétrique. Deux volets sont abordés dans ce contexte. Le premier, dans le cas de l'estimation de la moyenne conditionnelle, a permis de développer une famille d'estimateurs récursifs à noyau, présentant l'avantage d'être calculés de façon itérative. Le second développe le cas de la régression sur quantiles via la méthode Support Vector Machine (SVM) qui est une méthode d'apprentissage performante basée sur la minimisation d'un risque pénalisé dans un espace de Hilbert à noyau reproduisant.

ABSTRACT

My research work focuses on functional data analysis. This branch of statistics studies models when data can be considered as curves. Indeed, with competitive technology, some data are collected on sharp measure grids. Moreover, these data frequently come from the observation of continuous phenomena (for instance temperature measurements, growth measurements, ...), and it is natural to deal with these data as functions (of time, space) more than vectors of measure points. We consider that we observe some realizations of random variables valued in some functional space.

My researches mainly focused on regression models involving functional data. The starting point is the simplest model, known as the functional linear model, in which the variable of interest is real-valued, and the explanatory variable is valued in a functional space. I got interested in some extensions of this model.

The first extension considers missing data in the observations. First, I looked at the problem of imputing missing data in the response variable, in order to reconstruct a full dataset. This full dataset will allow a future prediction when a new observation of the covariate comes. Some perspectives of this work can be considered, for example taking into account missing data for the covariate as well as the response, or considering a response also valued in a functional space.

The second extension of my research work is a study where the response variable is valued in a functional space. This work introduces an estimation method of the functional parameter of the model and gives some asymptotic results for the prediction error. In the same time, I got interested in a sub-model considering (for example with variables measured over time) the case where the response at a certain time is explained by the past of the covariate before this time.

The third extension is the nonparametric one. Two topics are addressed in this context. The first one, in the case of the conditional mean estimation, allows to develop a recursive kernel estimator, showing the advantage to be computed iteratively. The second one develops the case of the quantile regression framework via the Support Vector Machine (SVM) method, which is a powerful learning method based on the minimization of a penalized risk in a reproducing kernel hilbert space.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction générale

Ce mémoire synthétise les travaux de recherche que j'ai pu mener depuis ma thèse. Le cadre de mon activité de recherche s'inscrit dans le domaine de la statistique fonctionnelle. L'idée fondamentale de cette branche de la statistique est de considérer des observations comme des courbes appartenant à un certain espace de fonctions (par exemple, des courbes de températures, des courbes de croissance au cours du temps, des images, ...). Une motivation principale de l'étude de ce type de données vient au départ des applications. En effet, les champs dans lesquels la statistique fonctionnelle peut potentiellement être utilisée sont nombreux : climatologie, linguistique, ... À titre d'exemple, la figure 1.1 représente 30 courbes journalières de la vitesse du vent (en mètres par seconde) mesurées par une éolienne à 99 mètres d'altitude en un certain lieu. Les mesures sont faites toutes les 10 minutes, soit 144 points de mesure par courbe.

Cette vision des données est le paradigme fondamental adopté dans l'étude de données fonctionnelles : la courbe est observée en pratique de manière discrétisée sur une grille d'observation, mais le phénomène sous-jacent étudié est un objet appartenant à un espace de fonctions. Les données de la figure 1.1 sont une matrice de taille 30×144, mais appliquer des méthodes de statistique multivariée sur ce type de données n'aurait pas de sens ici, au moins pour deux raisons :

- les points de mesure ne sont pas nécessairement les mêmes pour toutes les courbes,
- les points de mesure des courbes sont en général très nombreux, en raison des progrès pour acquérir à l'heure actuelle de grandes masses de données, et les problèmes à traiter sont en grande, voire très grande dimension, ce qui met en défaut les outils classiques de statistique multivariée.

Depuis la fin des années 1990, la statistique fonctionnelle n'a cessé de se développer, et les articles de recherche sur ce type de données se sont multipliés ces quinze dernières années. Plusieurs ouvrages fondateurs sont venus assoir les connaissances dans ce domaine, en faire une revue complète est utopique, néanmoins, nous présentons les principaux travaux dans ce qui suit.

La monographie Ramsay et Silverman (2005) est une contribution parmi les plus notables. Les problèmes préliminaires de collecte de données fonctionnelles y sont abordés, notamment les aspects de lissage, d'interpolation et de recalage de courbes. Puis, des méthodes exploratoires, bien connues en statistique multivariée, comme l'analyse

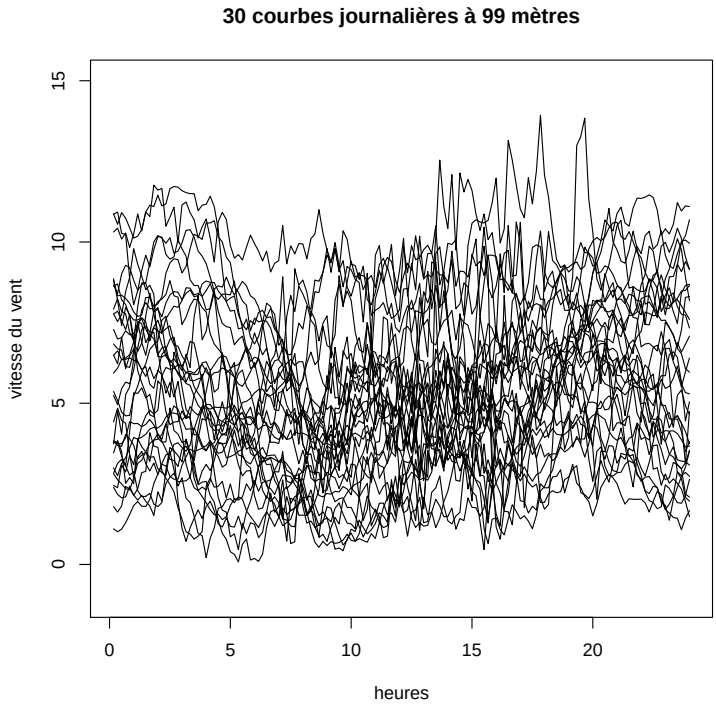

Figure 1.1 Représentation de 30 courbes journalières de la vitesse du vent (en mètres par seconde) à 99 mètres d'altitude.

en composantes principales ou l'analyse de corrélation canonique, sont présentées dans le contexte fonctionnel. Ensuite, une part importante est faite à la présentation de divers modèles de régression avec variables fonctionnelles. En fin d'ouvrage, des éléments complémentaires sont envisagées, par exemple la prise en compte de l'utilisation des dérivées des courbes afin de mieux comprendre l'information qu'elles contiennent.

En complément de cet ouvrage, une autre monographie Ramsay et Silverman (2002) a permis de présenter divers exemples d'applications où les données fonctionnelles apparaissent, laissant entrevoir la quantité impressionnante de champs d'applications, parmi lesquels on peut citer

- l'économie (indice de production de denrées périssables au cours du temps aux Etats-Unis),
- la biologie humaine (courbes de croissance d'enfants),
- la graphologie (étude de l'écriture de caractères),
- ...

D'un point de vue de l'implémentation, des fonctions R et Matlab sont présentées dans l'ouvrage Ramsay et al. (2009). Les auteurs définissent notamment des objets fonctionnels, utilisant l'idée que l'expression d'une courbe peut se résumer à la connaissance des coefficients de cette courbe dans une base de fonctions (Fourier, splines, ondelettes, ...).

Plus récemment, des ouvrages ont aussi creusé des aspects plus théoriques des données fonctionnelles, tout en gardant les ponts avec les applications, comme la monographie de Horváth et Kokoszka (2012) qui reprend les principaux modèles et méthodes en analyse de données fonctionnelles (analyse en composantes principales, analyse de corrélation canonique, modèles linéaires) pour des données indépendantes ou dépendantes. Zhang (2014) accorde un focus particulier à l'analyse de variance pour données fonctionnelles. Après avoir fait des rappels sur les données fonctionnelles et sur la reconstruction de courbes, il présente le problème du test de l'égalité de courbes moyennes de deux processus stochastiques. Le lien est fait avec des modèles linéaires fonctionnels avec réponses fonctionnelles qui se ramènent à une analyse de variance fonctionnelle. Récemment, Hsing et Eubank (2015) ont fait une synthèse de ces méthodes classiques (analyse en composantes principales, analyse de corrélation canonique, modèles linéaires, analyse de variance) en donnant une vision plus centrée sur la théorie des opérateurs, notamment les opérateurs compacts, en lien avec l'opérateur de covariance d'une variable aléatoire fonctionnelle. En particulier, le modèle linéaire fonctionnel est abordé sous l'angle de la régularisation et d'un problème de type moindres carrés pénalisé.

D'autres ouvrages, abordant des aspects plus spécifiques de l'analyse des données fonctionnelles ont également été écrits. Parmi les principaux, on peut citer la monographie Ferraty et Vieu (2006) qui explore intensivement les aspects non-paramétriques

sur des modèles fonctionnels. Shi et Choi (2011) adoptent un point de vue bayésien en étudiant la régression non linéaire bayésienne avec des priors processus gaussiens.

En parallèle de tous ces travaux, plusieurs ouvrages collectifs ont également été rédigés, permettant de montrer la grande variété des sujets abordés en analyse de données fonctionnelles. On peut citer à titre d'exemple Ferraty (2011).

Mes travaux de recherches se sont focalisés sur un thème central en statistique, celui de la régression. L'idée d'expliquer une variable d'intérêt à l'aide d'une ou plusieurs variables explicatives est très ancienne, remontant au début du XIX$^{\text{ème}}$ siècle (Legendre (1805)). D'innombrables travaux ont suivi, mais c'est dans les années 1990, au moment où le concept de données fonctionnelles émerge, que l'idée de régression sur variable fonctionnelle apparaît, et en particulier le modèle linéaire fonctionnel (voir Cardot et al. (1999)), dans sa version la plus simple : la variable réponse est réelle, la variable explicative est fonctionnelle, et la relation entre les deux est linéaire. En d'autres termes, si Y désigne la variable réponse, X désigne la variable explicative, alors on considère le modèle

$$Y = \langle \theta, X \rangle + \varepsilon, \tag{1.1}$$

où θ est la fonction inconnue du modèle et ε est une variable aléatoire centrée représentant l'erreur de modèle, avec variance finie $\mathbb{E}(\varepsilon^2) = \sigma_\varepsilon^2$. La variable explicative fonctionnelle X est à valeurs dans un espace H de fonctions, muni de son produit scalaire $\langle ., . \rangle$ et sa norme associée $\|.\|$. Par exemple, H peut être l'espace $L^2([a,b])$ des fonctions de carré intégrable définies sur un intervalle $I = [a,b]$ et le produit scalaire correspondant est défini par $\langle f, g \rangle = \int_a^b f(t)g(t)\,\mathrm{d}t$ pour toutes fonctions $f, g \in L^2([a,b])$. Sans perte de généralité, nous considérons notre travail avec des fonctions définies sur l'intervalle $I = [0,1]$. De plus, nous supposons que X et ε sont indépendantes. Ce modèle le plus simple de régression fonctionnelle peut être généralisé dans de nombreuses directions, dont j'ai exploré certaines dans mes travaux de recherche.

Une première façon de généraliser le modèle linéaire fonctionnel consiste à considérer la problématique des données manquantes. En effet, s'il est plus simple au départ de considérer un modèle pour lequel les données seront complètement observées, en pratique, de nombreuses situations peuvent faire apparaître des données manquantes. C'est autour de cette problématique que sera développé le chapitre 2. Dans ce travail, nous avons considéré des données manquantes sur la variable réponse, et construit une méthode d'imputation de ces données manquantes afin de reconstituer le jeu de données. Une fois le jeu de données initial reconstitué, il est possible d'estimer le paramètre fonctionnel du modèle, et de prédire une valeur pour la variable réponse lorsqu'une nouvelle observation X arrive dans l'échantillon.

Une autre généralisation possible du modèle linéaire fonctionnel est de considérer que la variable réponse est elle aussi fonctionnelle. Le modèle linéaire fonctionnel avec

sortie fonctionnelle s'écrit alors

$$Y(t) = \langle \theta(.,t), X \rangle + \varepsilon(t). \tag{1.2}$$

La principale problématique (le fait que l'estimation de θ est un problème mal posé, relié au problème d'inversion de l'opérateur de covariance de X) reste la même que dans le cas où la variable réponse est réelle. Nous avons étudié l'estimation de θ et l'erreur de prévision sur la réponse, ce travail est développé dans le chapitre 3.

Toujours autour d'une variable réponse fonctionnelle, un modèle restreint consiste à considérer (notamment lorsque les variables sont fonctions du temps) que la valeur de la variable réponse en un certain instant est expliquée par le passé de la variable explicative avant cet instant. Nous avons considéré un modèle de convolution s'écrivant

$$Y(t) = \int_0^t \theta(s)\, X(t-s) ds + \varepsilon(t). \tag{1.3}$$

L'estimation de θ dans ce modèle a été réalisée en appliquant la transformée de Fourier à ce modèle, de façon à utiliser les propriétés de la transformée de Fourier, permettant notamment de transformer une convolution en produit. Ce travail est développé dans le chapitre 4.

Jusqu'alors, le modèle considéré a toujours été linéaire. Dans certaines situations, cette modélisation est trop restrictive et il est préférable de se tourner vers une modélisation non-paramétrique. Le modèle considéré s'écrit alors

$$Y = \Psi(X) + \varepsilon, \tag{1.4}$$

où Ψ est un opérateur de L^2 dans $\mathbb{R}$ à estimer. L'estimation à noyau de Ψ est largement développée dans Ferraty et Vieu (2006). Notre travail, présenté dans le chapitre 5, a permis de développer une famille d'estimateurs récursifs à noyau, présentant l'avantage d'être calculés de façon itérative, cet avantage étant important lorsque de nouvelles données arrivent, l'estimateur devant être mis à jour régulièrement.

Pour finir, les différents modèles considérés précédemment sont reliés à l'estimation de la moyenne conditionnelle $\mathbb{E}(Y|X)$. Or, dans certaines situations, la moyenne conditionnelle n'est pas forcément la plus adaptée (présence de données aberrantes, données asymétriques), et les quantiles conditionnels sont une alternative intéressante. Dans le chapitre 6, nous considérons le modèle non-paramétrique

$$Y = \Psi_\tau(X) + \varepsilon_\tau, \tag{1.5}$$

où Ψ_τ est l'opérateur quantile à estimer et le bruit ε_τ vérifie $\mathbb{P}(\varepsilon_\tau \leq 0|X) = \tau$, $\tau \in]0;1[$ étant l'ordre du quantile. L'estimation de Ψ_τ est réalisée à l'aide de la méthode Support Vector Machine (SVM) qui est une méthode d'apprentissage performante basée sur la minimisation d'un risque pénalisé dans un espace de Hilbert à noyau reproduisant.

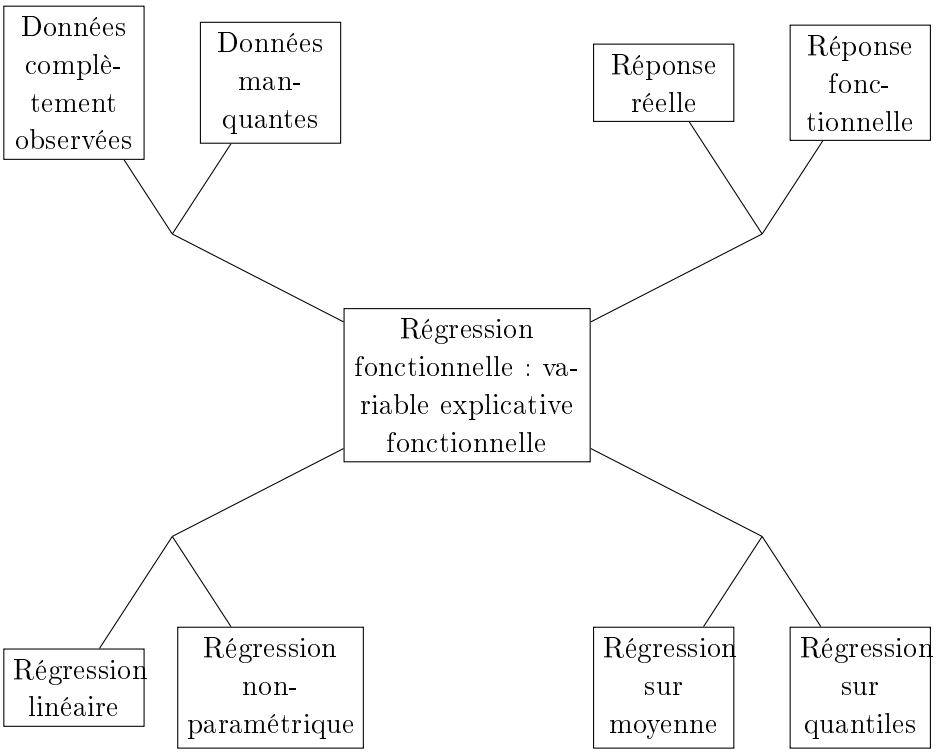

Figure 1.2 Différents modèles étudiés en régression fonctionnelle.

Une synthèse des différents modèles de régression fonctionnelle sur lesquels j'ai travaillé se trouve sur la Figure 1.2.

Bibliographie

H. Cardot, F. Ferraty et P. Sarda. Functional linear model. *Journal of Statistics and Probability Letters*, 45 :11–22, 1999.

F. Ferraty. Recent advances in functional data analysis and related topics. In *Contribution to Statistics F. Ferraty, (Eds.)*, pages 305–490. Physica-Verlag, Springer Heidelberg, 2011.

F. Ferraty et P. Vieu. *Nonparametric functional data analysis : Theory and practice.* Springer-Verlag, New York, 2006.

L. Horváth et P. Kokoszka. *Inference for Functional Data with Applications.* Springer-Verlag, New York, 2012.

T. Hsing et R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators.* Wiley series in probability and statistics, John Wiley & Sons, Chichester, 2015.

A.-M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes.* F. Didot, 1805.

J.O. Ramsay et B.W. Silverman. *Applied Functional Data Analysis.* Springer-Verlag, New York, 2002.

J.O. Ramsay et B.W. Silverman. *Functional Data Analysis (Second edition).* Springer-Verlag, New York, 2005.

J.O. Ramsay, G. Hooker et S. Graves. *Functional Data Analysis with R and Matlab.* Springer, New York, 2009.

J.Q. Shi et T. Choi. *Gaussian Process Regression Analysis for Functional Data.* Chapman and Hall, CRC, 2011.

J.T. Zhang. *Analysis of variance for functional data.* Monographs on statistics and applied probability, CRC Press, Boca Raton, 2014.

Chapitre 2

Imputation par régression dans le modèle linéaire fonctionnel avec réponses réelles manquantes

Contents

Les résultats présentés dans ce chapitre sont tirés de Crambes et Henchiri (2019).

2.1 Introduction

Dans ce chapitre, nous nous intéressons au modèle linéaire fonctionnel (voir Ramsay et Silverman (2005))

$$Y = \langle \theta, X \rangle + \varepsilon, \tag{2.1}$$

où θ est la fonction inconnue du modèle, Y est la variable d'intérêt à valeurs réelles, ε est une variable aléatoire centrée représentant l'erreur de modèle, avec variance finie $\mathbb{E}(\varepsilon^2) = \sigma_\varepsilon^2$, et X est la variable explicative fonctionnelle, indépendante de ε.

En analyse de données fonctionnelles, beaucoup de travaux concernent des cas où les données sont complètement observées, ce qui peut ne pas être le cas dans un grand nombre d'applications (par exemple, des cas de données de survie, lorsqu'un appareil de mesure tombe en panne, ...). Pour cette raison, nous allons nous focaliser sur le problème de données manquantes (voir Little et Rubin (2002); Graham (2012) pour un panorama sur les données manquantes dans un cadre multivarié). Ce sujet a été largement étudié, en particulier la façon d'imputer des données manquantes, et sur la précision de cette imputation vis-à-vis du type de données manquantes. En effet, les données manquantes peuvent être de type MCAR (Missing Completely At Random), MAR (Missing At Random), ou MNAR (Missing Not At Random). Beaucoup de travaux traitent de cette problématique lorsque les données sont multivariées, en revanche les travaux sont beaucoup plus marginaux pour les données fonctionnelles. Dans le cadre MAR, He et al. (2011) ont développé une méthode d'imputation multiple pour une réponse longitudinale dans une modèle fonctionnel à effets mixtes. De plus, Ferraty et al. (2013) ont considéré deux types d'estimation de la moyenne d'un salaire (variable d'intérêt réelle) dans un modèle de régression non paramétrique où la variable explicative fonctionnelle est complètement observée et la réponse est manquante. Également dans le cadre d'une réponse réelle et d'une variable explicative fonctionnelle, Febrero-Bande et al. (2019) considèrent une méthode d'imputation pour les données manquantes sur la variable réponse. Il s'agit du travail le plus proche du nôtre, bien que n'abordant pas les aspects théoriques de la méthode d'imputation. Notre travail propose des résultats sur ce plan. Dans le cadre MCAR, Preda et al. (2010) ont adapté la méthodologie de l'algorithme NIPALS (Nonlinear Iterative Partial Least Squares) pour proposer une méthode d'imputation pour des données manquantes qui affectent des variables explicatives fonctionnelles. Dans le cadre MNAR, Bugni (2012) étudie un test de spécification pour données fonctionnelles en présence d'observations manquantes. Enfin, Chiou et al. (2014) proposent une approche non-paramétrique pour l'imputation de données manquantes et la détection d'outliers sur des données fonctionnelles.

Il est important de bien faire la distinction entre un problème de données manquantes et un problème de prévision. En effet, le méchanisme de données manquantes implique une variable aléatoire δ (qui indique si la donnée est observée ou manquante) qui joue un rôle crucial dans le comportement de l'estimateur. Par exemple, cette variable δ et la variable explicative X sont dépendantes dans le cadre MAR. Ce point est également soulevé dans Ferraty et al. (2013). Dans notre travail, nous considérons une méthode d'imputation, basée sur les individus complètement observés, pour remplacer les valeurs manquantes dans la réponse du modèle linéaire fonctionnel. Une fois que la base de données est complète, nous pouvons estimer la fonction inconnue θ du modèle avec l'échantillon complété. Cet estimateur peut ensuite être utilisé pour prévoir de nouvelles valeurs de la réponse sur un échantillon test. Notre apport est en premier lieu théorique, en fournissant des vitesses de convergence des erreurs en moyenne quadratique pour les valeurs imputées, ainsi que sur les valeurs prédites sur un échantillon test utilisant l'échantillon reconstitué. Le comportement de la méthode en pratique sera illustré sur simulations.

2.2 Imputation d'une valeur manquante dans la réponse

2.2.1 Régression fonctionnelle sur composantes principales

Considérons un échantillon $(X_i, Y_i)_{i=1,\dots,n}$ indépendant et identiquement distribué, de même loi que le couple (X, Y). Une estimation de θ basée sur l'analyse en composantes principales des courbes $X_1, \dots, X_n$ a été étudiée dans de nombreux articles, voir par exemple Cardot et al. (1999). Nous rappelons ci-dessous la construction de cet estimateur. Considérons l'opérateur de covariance de X défini sous la condition $\mathbb{E}\left(\|X\|^2\right) < +\infty$ (supposée satisfaite dans tout ce qui suit) par

$$\Gamma u = \mathbb{E}\Big(\langle X, u\rangle X\Big),$$

pour tout $u \in H$ et sa version empirique

$$\widehat{\Gamma}_n u = \frac{1}{n}\sum_{i=1}^{n}\langle X_i, u\rangle X_i,$$

nous appelons $(\lambda_j)_{j\geq 1}$ $\Big(\text{resp. } \left(\widehat{\lambda}_j\right)_{j\geq 1}\Big)$ la suite de valeurs propres de Γ $\Big(\text{resp. } \widehat{\Gamma}_n\Big)$ et $(v_j)_{j\geq 1}$ $\Big(\text{resp. } (\widehat{v}_j)_{j\geq 1}\Big)$ la suite de fonctions propres de Γ $\Big(\text{resp. } \widehat{\Gamma}_n\Big)$. L'identifiabilité du modèle (2.1) est assurée sous l'hypothèse que $\lambda_1 > \lambda_2 > \ldots > 0$ (voir Cardot et al. (1999)). De plus, en supposant que $\widehat{\lambda}_{k_n} > 0$ pour un certain nombre entier k_n dépendant de n, l'estimateur de θ est défini par

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i, \widehat{v}_j\rangle Y_i}{\widehat{\lambda}_j}\,\widehat{v}_j. \tag{2.2}$$

Un résultat de consistence de cet estimateur est donné dans Cardot et al. (1999), des résultats plus récents peuvent être trouvés dans Cai et Hall (2006); Hall et Horowitz (2007).

2.2.2 Point de vue opératoriel

Notons dans cette sous-section que le modèle (2.1) peut-être abordé d'un point de vue opératoriel. Nous pouvons écrire le modèle

$$Y = \Theta X + \varepsilon, \tag{2.3}$$

où $\Theta : H \longrightarrow \mathbb{R}$ est un opérateur linéaire continu défini par $\Theta u = \langle \theta, u\rangle$ pour toute fonction $u \in H$. Considérons $\widehat{\Delta}_n$ l'opérateur de covariance croisée défini par $\widehat{\Delta}_n u = \frac{1}{n}\sum_{i=1}^{n}\langle X_i, u\rangle Y_i$, pour tout $u \in H$. Alors, l'estimateur $\widehat{\Theta}$ de Θ, vérifiant $\widehat{\Theta} = \langle \widehat{\theta}, .\rangle$, est donné par

$$\widehat{\Theta} = \widehat{\Pi}_{k_n}\widehat{\Delta}_n\left(\widehat{\Pi}_{k_n}\widehat{\Gamma}_n\right)^{-1}, \tag{2.4}$$

où $\widehat{\Pi}_{k_n}$ est l'opérateur de projection sur le sous-espace $\mathrm{span}(\widehat{v}_1, \ldots, \widehat{v}_{k_n})$.

2.2.3 Principe d'imputation

Présentons maintenant le contexte de données manquantes. Il peut y avoir de nombreuses raisons pour lesquelles des données manquantes peuvent apparaître : panne sur un appareil de mesure, une personne qui ne souhaite pas répondre à une question dans une enquête, ... Nous considérons que certaines observations $Y_1, \ldots, Y_n$ ne sont pas disponibles. Nous définissons la variable aléatoire réelle δ et nous considérons l'échantillon $(\delta_i)_{i=1,\ldots,n}$ tel que $\delta_i = 1$ si la valeur Y_i est observée et $\delta_i = 0$ si la valeur Y_i est manquante, pour tout $i = 1, \ldots, n$. Les données sont

$$\{(Y_i, \delta_i, X_i)\}_{i=1}^{n}.$$

Nous considérons que les données manquantes sont MAR. L'hypothèse MAR revient à dire que δ et Y sont indépendantes conditionnellement à X. En d'autres termes,

$$\mathbb{P}\left(\delta = 1 \mid X, Y\right) = \mathbb{P}\left(\delta = 1 \mid X\right). \tag{2.5}$$

Notons que l'hypothèse MAR est plus faible que l'hypothèse MCAR (pour laquelle $\mathbb{P}(\delta = 1 \mid X, Y) = \mathbb{P}(\delta = 1)$), puisqu'elle permet que le fait que les données soient manquantes dépendent potentiellement de la covariable, ce qui peut être une hypothèse raisonnable dans plusieurs cas pratiques. En conséquence de cette hypothèse MAR, la variable δ (le fait que l'observation soit manquante ou observée) est indépendante de l'erreur ε du modèle. Dans la suite, le nombre de données manquantes dans l'échantillon est noté

$$m_n = \sum_{i=1}^{n} \mathbb{1}_{\{\delta_i = 0\}}. \tag{2.6}$$

Ainsi, pour imputer une donnée manquante, disons Y_ℓ (où ℓ est un nombre entier donné entre 1 et n), une façon simple est de considérer l'analyse des cas complets (voir par exemple Little et Rubin (2002); Cheng (1994); Wang et al. (2004); Mojirsheibani (2007); Buuren (2012)). Cette méthode d'imputation par régression utilise les paires de données observées pour définir l'estimateur du coefficient fonctionnel du modèle. Plus précisément, nous définissons

$$Y_{\ell,imp} = \frac{1}{n - m_n} \sum_{\substack{i=1 \\ i \neq \ell}}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i, \widehat{v}_j \rangle \langle X_\ell, \widehat{v}_j \rangle \delta_i Y_i}{\widehat{\lambda}_j}. \tag{2.7}$$

D'un point de vue opératoriel, l'imputation de la valeur manquante Y_ℓ revient à

$$Y_{\ell,imp} = \widehat{\Pi}_{k_n,obs} \widehat{\Delta}_{n,obs} \left(\widehat{\Pi}_{k_n,obs} \widehat{\Gamma}_{n,obs} \right)^{-1} X_\ell, \tag{2.8}$$

où $\widehat{\Gamma}_{n,obs} = \frac{1}{n-m_n} \sum_{i=1}^{n} \langle X_i, . \rangle \delta_i X_i$, $\widehat{\Delta}_{n,obs} = \frac{1}{n-m_n} \sum_{i=1}^{n} \langle X_i, . \rangle \delta_i Y_i$ et $\widehat{\Pi}_{k_n,obs}$ est l'opérateur de projection sur le sous-espace $\mathrm{span}(\widehat{v}_{1,obs}, \ldots, \widehat{v}_{k_n,obs})$ où $\widehat{v}_{1,obs}, \ldots, \widehat{v}_{k_n,obs}$ sont les k_n premières fonctions propres de l'opérateur de covariance $\widehat{\Gamma}_{n,obs}$.

2.2.4 Estimation de θ et prévision

Une fois que la base de données est reconstruite par imputation des données manquantes, nous pouvons utiliser la base de données complète pour estimer le coefficient fonctionnel θ du modèle, directement à partir de (2.2) (voir Chu et Cheng (1995)), c'est-à-dire

$$\widetilde{\theta} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i, \widehat{v}_j \rangle Y_i^\star}{\widehat{\lambda}_j} \widehat{v}_j, \tag{2.9}$$

où $Y_i^\star = Y_i \delta_i + Y_{i,imp}(1 - \delta_i)$ pour tout $i = 1, \ldots, n$. Ainsi, cet estimateur de θ peut être utilisé pour faire de la prévision de nouvelles valeurs de la réponse Y sur un échantillon

test. Si X_{new} est une nouvelle courbe arrivant dans l'échantillon, la réponse prédite correspondante est

$$\widehat{Y}_{new} = \langle X_{new}, \widetilde{\theta} \rangle = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i, \widehat{v}_j \rangle \langle X_{new}, \widehat{v}_j \rangle Y_i^{\star}}{\widehat{\lambda}_j}. \tag{2.10}$$

2.3 Résultats théoriques

Dans cette section, nous présentons les principaux résultats obtenus. Les preuves de ces résultats sont donnés dans Crambes et Henchiri (2019). Pour cela, nous considérons les hypothèses suivantes.

$(A.1)$ Nous supposons qu'il existe une fonction convexe λ telle que $\lambda(j) = \lambda_j$ pour tout $j \geq 1$ qui interpole de façon continue les valeurs propres λ_j entre j et $j+1$.

$(A.2)$ Il existe une constante C strictement positive telle que

$$\mathbb{E}\left(\|X\|^4\right) \leq C.$$

Ces hypothèses sont relativement classiques dans notre contexte. L'hypothèse $(A.1)$ est semblable à une hypothèse de Crambes et Mas (2013). Il s'agit d'une condition relativement faible qui permet de considérer une large classe de décroissances de valeurs propres pour l'opérateur de covariance Γ, par exemple une décroissance polynomiale ou une décroissance exponentielle. L'hypothèse $(A.2)$ est vérifiée pour de nombreux processus X (processus Gaussiens, processus bornés, ...) et cette hypothèse est également présente dans Crambes et Mas (2013).

2.3.1 Résultat sur une valeur imputée et sur l'erreur globale

Theorème 2.1. *Si $(A.1)$ et $(A.2)$ sont vérifiées et si $\lambda_{k_n} k_n$ tend vers zéro quand n tend vers l'infini, nous avons*

$$\mathbb{E}\Big(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\Big)^2 = \sum_{j=k_n+1}^{+\infty} \Big(\Theta \Gamma^{1/2} v_j\Big)^2 + \frac{\sigma_\varepsilon^2 k_n}{n - m_n} + o\left(\frac{k_n}{n - m_n}\right),$$

et

$$\sum_{\ell=1}^{n} (1 - \delta_\ell) \mathbb{E}\Big(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\Big)^2 = m_n \sum_{j=k_n+1}^{+\infty} \Big(\Theta \Gamma^{1/2} v_j\Big)^2 + \frac{\sigma_\varepsilon^2 k_n m_n}{n - m_n} + o\left(\frac{k_n m_n}{n - m_n}\right).$$

2.3.2 Vitesses de convergence

Pour préciser la vitesse de convergence de l'erreur quadratique moyenne commise en remplaçant la vraie valeur $\langle \theta, X_\ell \rangle$ par la valeur imputée $Y_{\ell,imp}$, nous introduisons une notation. Pour toute fonction $\varphi : \mathbb{R}_+^\star \longrightarrow \mathbb{R}_+^\star$ pour tout nombre réel positif L, nous définissons

$$\mathcal{C}(\varphi, L) = \left\{ T : H \longrightarrow \mathbb{R} \ / \ \forall j \geq 1, Tv_j \leq L\sqrt{\varphi(j)} \right\}.$$

Nous pouvons noter que le fait que $\Theta\Gamma^{1/2}$ appartienne à $\mathcal{C}(\varphi, L)$ est satisfait dans des cas simples. Par exemple, considérons l'opérateur Θ exprimé dans la base des fonctions propres $(v_j)_{j\geq 1}$ tel que $\Theta u = \sum_{j=1}^{+\infty} \theta_j \langle v_j, u \rangle$ pour tout $u \in H$, avec θ_j qui tend vers zéro lorsque j tend vers l'infini. Alors il existe une borne L telle que $\theta_j \leq L$ pour tout $j \geq 1$ et $\Theta\Gamma^{1/2} v_j = \theta_j \sqrt{\lambda_j} \leq L\sqrt{\lambda_j}$.

Theorème 2.2. *Soit* $L = \left\| \Theta\Gamma^{1/2} \right\|_\infty$ *et* φ *la fonction définie par* $\varphi(j) = \frac{\left(\Theta\Gamma^{1/2} v_j\right)^2}{L^2}$ *pour tout* $j \geq 1$ *qui interpole de façon continue les valeurs* $\varphi(j)$ *entre* j *et* $j+1$. *Sous les hypothèses* $(A.1)$-$(A.2)$, *l'opérateur* $\Theta\Gamma^{1/2}$ *appartient à* $\mathcal{C}(\varphi, L)$ *et*

$$\mathbb{E}\left(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\right)^2 \underset{n\to+\infty}{\sim} 2\sigma_\varepsilon^2 \frac{k_n^\star}{n - m_n},$$

où $k_n^\star$ *est la solution de l'équation en* x

$$\int_x^{+\infty} \varphi(t)\, \mathrm{d}t = \frac{\sigma_\varepsilon^2}{L^2(n - m_n)} x. \tag{2.11}$$

Notons que cette équation correspond au compromis habituel que l'on fait entre le biais au carré $\sum_{j=k_n+1}^{+\infty} \left(\Theta\Gamma^{1/2} v_j\right)^2$ *(approximé par une intégrale) et la variance* $\frac{\sigma_\varepsilon^2 k_n}{n - m_n}$ *obtenus dans le Théorème 2.1. Ainsi,* $k_n^\star$ *est le nombre optimal de composantes principales vis-à-vis du critère de l'erreur quadratique moyenne.*

À nouveau, pour l'erreur quadratique moyenne sur toutes les valeurs imputées, nous avons

$$\sum_{\ell=1}^{n} (1 - \delta_\ell) \mathbb{E}\left(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\right)^2 \underset{n\to+\infty}{\sim} 2\sigma_\varepsilon^2 \frac{k_n^\star m_n}{n - m_n}.$$

Remarque 2.1. *Notons que l'équation (2.11) a une unique solution (le terme de gauche est décroissant en* x, *celui de droite est croissant en* x). *Cependant, la résolution explicite de cette équation est difficile en général à cause du calcul de* L. *Pour contourner ce problème, nous proposons une autre façon de choisir le nombre de composantes principales en pratique (voir la section 2.4 dans la suite).*

Le résultat du Théorème 2.2 est semblable à celui obtenu par Crambes et Mas (2013) (qui considèrent que la réponse, fonctionnelle, est complètement observée). La vitesse de convergence est affectée ici par le nombre m_n de données manquantes. Nous précisons le résultat de convergence dans les exemples suivants.

Exemple 2.1. *Nous considérons deux types de fonctions φ tels que $\varphi_{pol}(j) = C_\alpha j^{-(2+\alpha)}$ et $\varphi_{exp}(j) = D_\alpha \exp(-\alpha j)$ où C_α et D_α sont des constantes positives et $\alpha > 0$. Alors, la solution de l'équation (2.11) vérifie*

$$\begin{cases} k^\star_{n,pol} \underset{n\to+\infty}{\sim} \left(\frac{C_\alpha L^2}{(1+\alpha)\sigma_\varepsilon^2}\right)^{1/(2+\alpha)} n^{1/(2+\alpha)}, & si\ \varphi = \varphi_{pol}, \\ k^\star_{n,exp} \underset{n\to+\infty}{\lesssim} \frac{\log n}{\alpha}, & si\ \varphi = \varphi_{exp}. \end{cases}$$

Pour $\varphi = \varphi_{pol}$, le résultat du Théorème 2.2 devient

$$\mathbb{E}\left(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\right)^2 \underset{n\to+\infty}{\sim} 2\left(\sigma_\varepsilon^2\right)^{(1+\alpha)/(2+\alpha)} \left(\frac{C_\alpha L^2}{1+\alpha}\right)^{1/(2+\alpha)} \frac{n^{1/(2+\alpha)}}{n-m_n},$$

pour l'imputation d'une valeur et

$$\sum_{\ell=1}^{n}(1-\delta_\ell)\mathbb{E}\left(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\right)^2 \underset{n\to+\infty}{\sim} 2\left(\sigma_\varepsilon^2\right)^{(1+\alpha)/(2+\alpha)} \left(\frac{C_\alpha L^2}{1+\alpha}\right)^{1/(2+\alpha)} \frac{n^{1/(2+\alpha)} m_n}{n-m_n},$$

pour l'erreur sur toutes les valeurs imputées.

Pour $\varphi = \varphi_{exp}$, le résultat du Théorème 2.2 devient

$$\mathbb{E}\left(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\right)^2 \underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 \log n}{\alpha(n-m_n)},$$

pour l'imputation d'une valeur et

$$\sum_{\ell=1}^{n}(1-\delta_\ell)\mathbb{E}\left(Y_{\ell,imp} - \langle \theta, X_\ell \rangle\right)^2 \underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 m_n \log n}{\alpha(n-m_n)},$$

pour l'erreur sur toutes les valeurs imputées.

Exemple 2.2. *Il est possible de préciser les vitesses de convergence dans des cas plus spécifiques, en considérant trois niveaux de données manquantes : (i) le nombre m_n de données manquantes est négligeable devant la taille de l'échantillon n, soit $m_n = a_n n$ avec a_n qui tend vers zéro lorsque n tend vers l'infini, (ii) le nombre de données*

	$\varphi = \varphi_{\text{pol}}$	$\varphi = \varphi_{\exp}$
$m_n := a_n n = o(n)$	$\underset{n\to+\infty}{\sim} K_\alpha n^{-(1+\alpha)/(2+\alpha)}$	$\underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 \log n}{\alpha n}$
$m_n = \rho n$	$\underset{n\to+\infty}{\sim} K_\alpha (1-\rho)^{1/(2+\alpha)} n^{-(1+\alpha)/(2+\alpha)}$	$\underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 \log n}{\alpha(1-\rho)n}$
$u_n := n - m_n = o(n)$	$\underset{n\to+\infty}{\sim} K_\alpha u_n^{-(1+\alpha)/(2+\alpha)}$	$\underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 \log n}{\alpha u_n}$

Table 2.1 Vitesses de convergence de l'erreur quadratique moyenne pour l'imputation d'une seule valeur, avec $K_\alpha := 2\left(\sigma_\varepsilon^2\right)^{(1+\alpha)/(2+\alpha)} \left(\frac{C_\alpha L^2}{1+\alpha}\right)^{1/(2+\alpha)}$.

	$\varphi = \varphi_{\text{pol}}$	$\varphi = \varphi_{\exp}$
$m_n := a_n n = o(n)$	$\underset{n\to+\infty}{\sim} K_\alpha a_n n^{1/(2+\alpha)}$	$\underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 a_n \log n}{\alpha}$
$m_n = \rho n$	$\underset{n\to+\infty}{\sim} K_\alpha \rho(1-\rho)^{1/(2+\alpha)} n^{1/(2+\alpha)}$	$\underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 \rho \log n}{\alpha(1-\rho)}$
$u_n := n - m_n = o(n)$	$\underset{n\to+\infty}{\sim} K_\alpha n u_n^{-(1+\alpha)/(2+\alpha)}$	$\underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 n \log n}{\alpha u_n}$

Table 2.2 Vitesses de convergence de l'erreur quadratique moyenne pour toutes les valeurs imputées, où $K_\alpha := 2\left(\sigma_\varepsilon^2\right)^{(1+\alpha)/(2+\alpha)} \left(\frac{C_\alpha L^2}{1+\alpha}\right)^{1/(2+\alpha)}$.

manquantes est proportionnel à la taille de l'échantillon, soit $m_n = \rho n$ avec $0 < \rho < 1$, et (iii) le nombre de valeurs observées est négligeable devant la taille de l'échantillon, soit $u_n := n - m_n = o(n)$. Nous pouvons résumer les vitesses de convergence pour l'erreur quadratique moyenne d'une seule valeur imputée (Table 2.1) et pour l'erreur sur toutes les valeurs imputées (Table 2.2).

Nous pouvons voir que les données manquantes n'affectent pas la vitesse de convergence de l'erreur quadratique moyenne sur une seule valeur imputée lorsqu'il n'y a pas trop de données manquantes ($m_n = o(n)$ ou $m_n = \rho n$). La vitesse $1/n^{(1+\alpha)/(2+\alpha)}$ correspond aux vitesses optimales usuelles dans ce contexte. La vitesse $\log n/\alpha n$ n'est pas exacte mais forcément précise puisque c'est une vitesse paramétrique à un terme logarithmique près. En revanche, la vitesse est davantage affectée lorsque le nombre de données manquantes est important ($m_n \sim n$). Pour l'erreur sur toutes les valeurs imputées, lorsqu'il n'y a pas trop de données manquantes ($m_n = o(n)$), le nombre de données manquantes joue un rôle crucial dans la vitesse, puisque la convergence dépend du fait que $a_n n^{1/(2+\alpha)}$ ou $a_n \log n$ tend vers zéro lorsque n tend vers l'infini. Dans les autres cas ($m_n = \rho n$ ou $m_n \sim n$), les données manquantes affectent la convergence de l'erreur sur toutes les valeurs imputées, celle-ci ne pouvant pas converger vers zéro.

2.3.3 Résultat sur l'estimation de θ et la prévision

Nous donnons ici un résultat qui montre que, à la condition qu'il n'y ait pas trop de données manquantes, la vitesse de convergence de l'erreur de prédiction en moyenne quadratique pour une nouvelle observation reste la même que dans le cas où il n'y a pas de données manquantes.

Theorème 2.3. *Sous les hypothèses du Théorème 2.1, si nous supposons de plus que* $m_n = o(n)$ *et* $m_n^2 k_n = O(n)$*, alors*

$$\mathbb{E}\Big(Y_{new} - \langle \theta, X_{new} \rangle\Big)^2 = \sum_{j=k_n+1}^{+\infty} \Big(\Theta \Gamma^{1/2} v_j\Big)^2 + O\left(\frac{k_n}{n}\right).$$

2.4 Applications

2.4.1 Mise en œuvre

En pratique, nous avons utilisé une version lissée de l'estimateur (2.2) basé sur le travail de Cardot et al. (2003) (régression lissée sur composantes principales). Pour le lissage, nous considérons une base de fonctions splines de régression avec les paramètres suivants : le nombre κ de nœuds des fonctions splines, le degré q des fonctions splines et l'ordre m de dérivation. Remarquons que, sous des conditions appropriées, les résultats théoriques développés dans la section 2.3 s'appliqueront aussi à cette version lissée de l'estimateur. Par exemple, nous supposons que l'estimateur $\widetilde{\theta}$ est r' fois dérivable pour un certain entier r' et $\widetilde{\theta}^{(r')}$ vérifie, pour un certain $\nu \in]0,1]$

$$\left|\widetilde{\theta}^{(r')}(t_1) - \widetilde{\theta}^{(r')}(t_2)\right| \leq C \left|t_1 - t_2\right|^{\nu},$$

pour tous $t_1, t_2 \in [0,1]$. Si nous notons $r = r' + \nu$ et si nous supposons que le degré q des fonctions splines est tel que $q \geq r$, alors

$$\sup_{t\in[0,1]} \left|\widetilde{\theta}(t) - S_{\kappa,q}(\widetilde{\theta})(t)\right| = O\left(\kappa^{-r}\right),$$

où $S_{\kappa,q}(\widetilde{\theta})$ est l'approximation spline de $\widetilde{\theta}$ (voir de Boor (1978)).

2.4.2 Choix du nombre de composantes principales

Dans ce paragraphe, nous nous intéressons à la procédure pour choisir le nombre k_n de composantes principales. Comme dit dans la section 2.3, il n'est pas possible dans de nombreux cas d'utiliser la relation théorique (2.11) pour déterminer le k_n optimal. Pour contourner ce problème, nous nous sommes tournés vers des critères classiques dans ce contexte : la validation croisée (CV), la validation croisée K-fold (K-fold CV) ou encore

la validation croisée généralisée (GCV) et nous avons sélectionné le paramètre optimal k_n^* minimisant ces critères. Ceux-ci sont définis pour l'imputation par

$$\begin{aligned}
\mathrm{CV}(k_n) &= \frac{1}{n-m_n}\sum_{i=1}^{n}(\hat{Y}_i^{[-i]} - \langle\theta, X_i\rangle)^2\delta_i, \\
\text{K-fold CV}(k_n) &= \frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}|B_k|^{-1}\sum_{i\in B_k}(\hat{Y}_i^{[-B_k]} - \langle\theta, X_i\rangle)^2\delta_i, \\
\mathrm{GCV}(k_n) &= \frac{(n-m_n)\sum_{i=1}^{n}(\hat{Y}_i - \langle\theta, X_i\rangle)^2\delta_i}{((n-m_n)-k_n)^2},
\end{aligned}$$

et de façon analogue pour la prédiction

$$\begin{aligned}
\mathrm{CV}(k_n) &= \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i^{*[-i]} - \langle\theta, X_i\rangle)^2, \\
\text{K-fold CV}(k_n) &= \frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}|B_k|^{-1}\sum_{i\in B_k}(\hat{Y}_i^{*[-B_k]} - \langle\theta, X_i\rangle)^2, \\
\mathrm{GCV}(k_n) &= \frac{n\sum_{i=1}^{n}(\hat{Y}_i^{*} - \langle\theta, X_i\rangle)^2}{(n-k_n)^2},
\end{aligned}$$

où $\hat{Y}_i^{[-i]}$ et $\hat{Y}_i^{[-B_k]}$ signifient respectivement que la valeur de Y_i est prédite en utilisant tout l'échantillon sauf la $i^{\text{ème}}$ observation ou sauf l'ensemble des observations indexées dans B_k. Les définitions sont analogues pour $\hat{Y}_i^{*[-i]}$ et $\hat{Y}_i^{*[-B_k]}$. Pour le cas de la validation croisée K-fold, les données sont partitionnées en K sous-ensembles $B_1, \ldots, B_K$ de taille aussi proche que possible.

Ces différents critères se sont montrés relativement proches sur simulations, le critère GCV ayant finalement été retenu pour sa rapidité de temps de calcul.

2.4.3 Simulations

Nous présentons dans ce paragraphe une brève simulation dans le cadre MAR pour observer le comportement de notre méthode en pratique. Une étude plus complète est faite dans Crambes et Henchiri (2019). Le modèle considéré est le suivant :

$$Y = \int_0^1 \sin(4\pi t)X(t)\mathrm{d}t + \varepsilon,$$

où l'erreur ε est la Gaussienne $N(0; 0.2)$ et X est un mouvement Brownien standard sur $[0; 1]$. Les tailles d'échantillons simulés sont respectivement $n = 100, 300$ et 1200 pour

les échantillons d'apprentissage $(X_1, Y_1), \ldots, (X_n, Y_n)$ et $n_1 = 50, 150$ et 600 pour les échantillons de test $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+n_1}, Y_{n+n_1})$. Les trajectoires des courbes X_i sont discrétisées sur $p = 100$ points équidistants.

Pour contrôler le nombre de données manquantes dans les simulations dans le cadre MAR, nous simulons δ suivant une régression logistique fonctionnelle. La variable δ suit la loi de Bernoulli de paramètre $p(X)$ tel que

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \langle \alpha_0, X\rangle + c,$$

où $\alpha_0(t) = \sin(2\pi t)$ pour tout $t \in [0,1]$ et c est une constante permettant de considérer différents niveaux de données manquantes. Nous prenons $c = 2$ pour environ 12.5% de données manquantes, $c = 1$ pour environ 27.4% de données manquantes et $c = 0.2$ pour environ 44.9% de données manquantes.

Pour évaluer le comportement de l'imputation sur l'échantillon d'apprentissage et de la prévision sur l'échantillon test, nous avons réalisé $S = 500$ simulations. Les critères considérés pour évaluer la qualité de l'imputation sur l'échantillon d'apprentissage sont l'erreur quadratique moyenne

$$\overline{MSE} = \frac{1}{S}\sum_{j=1}^{S}\frac{1}{m_n}\sum_{\ell=1}^{n}\left(Y_{\ell,imp}^{j} - \langle \theta, X_\ell^j\rangle\right)^2(1-\delta_\ell),$$

et l'erreur quadratique relative moyenne (ratio respect to truth en anglais)

$$\overline{RT} = \frac{1}{S}\sum_{j=1}^{S}\frac{\displaystyle\sum_{\ell=1}^{n}\left(Y_{\ell,imp}^{j} - \langle \theta, X_\ell^j\rangle\right)^2(1-\delta_\ell)}{\displaystyle\sum_{\ell=1}^{n}\left(\varepsilon_\ell^j\right)^2(1-\delta_\ell)}.$$

En ce qui concerne la qualité de la prévision sur l'échantillon test, les critères considérés sont également l'erreur quadratique moyenne $\overline{MSE'}$ et l'erreur quadratique relative moyenne $\overline{RT'}$ calculés cette fois sur les n_1 observations de l'échantillon test.

Les résultats des simulations sont regroupés dans les tables 2.3 et 2.4. Nous constatons comme attendu que les erreurs diminuent lorsque la taille de l'échantillon augmente et les erreurs augmentent lorque le pourcentage de données manquantes augmente dans l'échantillon. Nous avons également réalisé des simulations dans le cadre MCAR, où nous avons constaté que les erreurs sont moins importantes que dans le cadre MAR pour un pourcentage de données manquantes équivalent.

	$n = 100$		
Données manquantes (%)			
Moyenne	12.520	27.420	44.882
Médiane	13	27	45
Écart-type	3.307	4.515	5.038
$\overline{MSE} \times 10^3$	2.3592	2.7845	3.2821
	(1.8375)	(2.0370)	(2.0679)
$\overline{RT} \times 10^2$	7.0001	7.5194	8.6148
	(6.6216)	(5.7701)	(5.7158)
	$n = 300$		
Données manquantes (%)			
Moyenne	12.433	27.456	45.209
Médiane	12.333	27.333	45.333
Écart-type	1.877	2.487	3.041
$\overline{MSE} \times 10^3$	0.8349	1.0048	1.3364
	(0.5728)	(0.6843)	(0.9037)
$\overline{RT} \times 10^2$	2.2327	2.5724	3.4547
	(1.5754)	(1.7245)	(2.3383)
	$n = 1200$		
Données manquantes (%)			
Moyenne	12.529	27.536	45.213
Médiane	12.500	27.500	45.250
Écart-type	0.934	1.280	1.355
$\overline{MSE} \times 10^3$	0.2326	0.2759	0.3521
	(0.1321)	(0.1519)	(0.2018)
$\overline{RT} \times 10^2$	0.5933	0.6962	0.8822
	(0.3492)	(0.3891)	(0.5036)

Table 2.3 Critères MSE et RT pour les valeurs imputées sur l'échantillon d'apprentissage, calculés sur $S = 500$ simulations.

2.5 Perspectives

2.5.1 Imputation multiple

Dans le travail présenté dans ce chapitre, nous nous sommes intéressés uniquement à l'imputation simple d'une donnée manquante. Il pourrait être intéressant de développer la méthodologie d'imputation multiple (voir Little et Rubin (2002), Buuren (2012)) dans ce contexte. L'imputation multiple est basée sur trois étapes : (i) l'imputation (on effectue une imputation des données manquantes, non pas une seule fois, mais m fois, en

	$n_1 = 50$		
Données manquantes (%)			
Moyenne	12.520	27.420	44.882
Médiane	13	27	45
Écart-type	3.307	4.515	5.038
$\overline{MSE'} \times 10^3$	2.3383	2.7173	3.1939
	(1.4987)	(1.8390)	(2.0391)
$\overline{RT'} \times 10^2$	5.9523	6.9769	8.2677
	(3.7338)	(4.9933)	(5.6516)
	$n_1 = 150$		
Données manquantes (%)			
Moyenne	12.433	27.456	45.209
Médiane	12.333	27.333	45.333
Écart-type	1.877	2.487	3.041
$\overline{MSE'} \times 10^3$	0.8453	0.9984	1.3046
	(0.5530)	(0.6729)	(0.8897)
$\overline{RT'} \times 10^2$	2.1534	2.5348	3.3255
	(1.3984)	(1.6629)	(2.2417)
	$n_1 = 600$		
Données manquantes (%)			
Moyenne	12.529	27.536	45.213
Médiane	12.500	27.500	45.250
Écart-type	0.934	1.280	1.355
$\overline{MSE'} \times 10^3$	0.2295	0.2746	0.3474
	(0.1282)	(0.1512)	(0.1982)
$\overline{RT'} \times 10^2$	0.5756	0.6887	0.8699
	(0.3165)	(0.3753)	(0.4888)

Table 2.4 Critères MSE' et RT' pour les prévisions sur l'échantillon test, calculés sur $S = 500$ simulations.

les tirant suivant une certaine distribution), (ii) l'analyse (les m échantillons complétés sont analysés par rapport à un certain objectif, par exemple construire un estimateur du coefficient inconnu dans le modèle), (iii) le groupement (on regroupe les résultats des m analyses pour en sortir un résultat final).

2.5.2 Modèle linéaire fonctionnel à sortie fonctionnelle

L'autre objectif pour poursuivre ce travail est de s'intéresser au modèle linéaire fonctionnel à sortie fonctionnelle qui s'écrit sous la forme

$$Y(t) = \langle \theta(.,t), X \rangle + \varepsilon(t),$$

pour tout $t \in I$. Dans ce modèle, la variable explicative et la variable à expliquer sont fonctionnelles. Certains travaux se sont intéressés à la complétion de courbes dont certaines parties sont manquantes, comme par exemple Kraus (2015), Kneip et Liebl (2019). Ces derniers pourraient être utilisés pour aborder le problème de l'estimation de θ dans le cadre où des données sont manquantes à la fois sur la variable d'entrée X et sur la variable de sortie Y. Cet axe de recherche correspond au début de la thèse (avril 2019) de Chayma Daayeb que je co-encadre. La procédure d'estimation qui a commencé à être développée est en deux étapes : (i) la reconstruction des parties manquantes sur les courbes explicatives, à l'aide de la méthodologie développée par Kneip et Liebl (2019), (ii) l'imputation des parties manquantes sur les courbes réponses en adaptant la méthodologie d'imputation présentée dans ce chapître au cadre où la réponse est une courbe.

2.5.3 Scores de propension

La probabilité $\pi(X) := \mathcal{P}(\delta = 1|X)$ d'observer Y sachant X est appelée score de propension (en anglais, propensity score). En utilisant une estimation $\widehat{\pi}_i$ des scores de propension π_i (voir par exemple Dubnicka (2009) qui introduit une estimation non-paramétrique à noyau), on peut imaginer définir l'imputation d'une donnée manquante Y_ℓ par

$$Y_{\ell,imp} = \frac{1}{n} \sum_{\substack{i=1 \\ i \neq \ell}}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i, \widehat{v}_j \rangle \langle X_\ell, \widehat{v}_j \rangle \delta_i Y_i}{\widehat{\pi}_i \widehat{\lambda}_j}.$$

Pour l'instant, le comportement de cet estimateur en pratique est assez instable numériquement. En ce sens, la qualité de l'estimation des scores de propension est cruciale. Ce travail et l'étude de cet estimateur reste un problème ouvert.

Bibliographie

F.A. Bugni. Specification test for missing functional data. *Econometric Theory*, 28 : 959–1002, 2012.

S. Van Buuren. *Flexible Imputation of Missing Data.* NJ : Chapman and Hall (CRC Press), Hoboken, 2012.

T.T. Cai et P. Hall. Prediction in functional linear regression. *The Annals of Statistics*, 34 :2159–2179, 2006.

H. Cardot, F. Ferraty et P. Sarda. Functional linear model. *Journal of Statistics and Probability Letters*, 45 :11–22, 1999.

H. Cardot, F. Ferraty et P. Sarda. Spline estimators for the functional linear model. *Journal of Statistica Sinica*, 13(3) :571–591, 2003.

P.E. Cheng. Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89 :81–87, 1994.

J-M. Chiou, Zhang Y-C., Chen W-H. et C-W. Chang. A functional data approach to missing value imputation and outlier detection for traffic flow data. *Transportmetrica B : Transport Dynamics*, 2 :106–129, 2014.

C.K. Chu et P.E. Cheng. Nonparametric regression estimation with missing data. *Journal of Statistical Planning and Inference*, 48 :85–99, 1995.

C. Crambes et Y. Henchiri. Regression imputation in the functional linear model with missing values in the response. *Journal of Statistical Planning and Inference*, 201 : 103–119, 2019.

C. Crambes et A. Mas. Asymptotics of prediction in functional linear regression with functional outputs. *Bernoulli*, 19 :2627–2651, 2013.

C. de Boor. *A Practical Guide to Splines.* Springer-Verlag, New York, 1978.

S. R. Dubnicka. Kernel density estimation with missing data and auxiliary variables. *Autralian and New Zealand Journal of Statistics*, 51 :247–270, 2009.

M. Febrero-Bande, P. Galeano et W. Gonzalez-Manteiga. Estimation, imputation and prediction for the functional linear model with scalar response with responses missing at random. *Computational Statistics and Data Analysis*, 131 :91–103, 2019.

F. Ferraty, M. Sued et P. Vieu. Mean estimation with data missing at random for functional covariables. *Statistics*, 47 :688–706, 2013.

J.W. Graham. *Missing data analysis and design.* Springer, New York, 2012.

P. Hall et J.L. Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1) :70–91, 2007.

Y. He, R. Yucel et T.E. Raghunathan. A functional multiple imputation approach to incomplete longitudinal data. *Statistics in medicine*, 30 :1137–1156, 2011.

A. Kneip et D. Liebl. On the optimal reconstruction of partially observed functional data. *Annals of Statistics*, 2019.

D. Kraus. Components and completion of partially observed functional data. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 77(4) :777–801, 2015.

R.J.A. Little et D.B. Rubin. *Statistical analysis with missing data (Second edition).* John Wiley, New York, 2002.

M. Mojirsheibani. Nonparametric curve estimation with missing data : A general empirical process approach. *Journal of Statistical Planning and Inference*, 137 :2733–2758, 2007.

C. Preda, G. Saporta et M.M.H. Hadj. The NIPALS algorithm for functional data. *Revue Roumaine de Mathématiques Pures et Appliquées*, 55 :315–326, 2010.

J.O. Ramsay et B.W. Silverman. *Functional Data Analysis (Second edition).* Springer-Verlag, New York, 2005.

Q. Wang, O. Linton et W. Härdle. Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association*, 99 :334–345, 2004.

Chapitre 3

Régression fonctionnelle sur composantes principales avec réponse fonctionnelle

Contents

Les résultats présentés dans ce chapitre sont tirés de Crambes et Mas (2013) et Crambes et al. (2016).

3.1 Introduction

Dans ce chapitre, nous présentons des travaux relatifs au modèle linéaire complètement fonctionnel (voir Ramsay et Silverman (2005)), c'est-à-dire lorsque la variable explicative et la variable réponse sont toutes deux fonctionnelles. Pout tout $t \in I$, ce modèle s'écrit dans sa version opératorielle

$$Y(t) = \Theta X(t) + \varepsilon(t), \tag{3.1}$$

et dans sa version fonctionnelle

$$Y(t) = \langle \theta(.,t), X \rangle + \varepsilon(t), \tag{3.2}$$

où l'erreur ε est indépendante de X.

Le modèle (3.2) a été le sujet de plusieurs études, comme par exemple Chiou et al. (2004) ou Yao et al. (2005), qui proposent une estimation du paramètre fonctionnel θ en utilisant les analyses en composantes principales fonctionnelles des courbes X et Y. Une des premières études sur ce modèle est due à Cuevas et al. (2002) qui ont considéré le cas d'un design fixe. Un estimateur spline du coefficient fonctionnel du modèle a été proposé par Antoch et al. (2008), tandis que Aguilera et al. (2008) introduisent un estimateur basé sur des ondelettes. D'autres travaux sont apparus plus récemment, par exemple Benatia et al. (2017) qui s'intéressent à une procédure d'estimation basée sur une régularisation de type Tikhonov, ou Imaizumi et Kato (2018) qui s'intéressent à l'estimation du paramètre fonctionnel θ dans le modèle (3.2) en utilisant l'analyse en composantes principales fonctionnelle des courbes explicatives.

Dans notre travail, l'objectif est focalisé sur la prédiction de la variable réponse Y. Dans un premier temps, nous allons présenter la procédure d'estimation de l'opérateur Θ du modèle (3.1) d'où nous déduisons la prédiction de la variable réponse de ce modèle. Les contributions de ce travail sont en premier lieu théoriques, avec une obtention de vitesses de convergence de l'erreur quadratique de prévision, optimales dans un certain sens. Dans un second temps, nous nous intéresserons à l'estimation de la variance du bruit, une quantité qui apparaît de façon cruciale dans nos résultats théoriques. Nous présentons une étude sur simulations en nous focalisant plus particulièrement sur le choix du nombre de composantes principales.

3.2 Estimation et prévision

Nous adoptons un point de vue basé sur la régression fonctionnelle sur composantes principales. Pour cela, de façon analogue au chapitre 2, nous considérons l'opérateur de covariance Γ de X

$$\Gamma = \mathbb{E}\left(\langle X, .\rangle X\right), \tag{3.3}$$

et l'opérateur de covariance croisée Δ entre X et Y

$$\Delta = \mathbb{E}\left(\langle X, .\rangle Y\right). \tag{3.4}$$

Comme X et ε sont indépendants, l'opérateur Θ vérifie l'équation de moments

$$\Delta = \Theta\Gamma. \tag{3.5}$$

Pour estimer l'opérateur Θ sur la base d'un échantillon $(X_i, Y_i)_{i=1,\dots,n}$, nous définissons les versions empiriques des opérateurs de covariance

$$\widehat{\Gamma}_n = \frac{1}{n}\sum_{i=1}^{n}\langle X_i, .\rangle X_i, \tag{3.6}$$

et de covariance croisée

$$\widehat{\Delta}_n = \frac{1}{n}\sum_{i=1}^{n}\langle X_i, .\rangle Y_i. \tag{3.7}$$

L'estimation de l'opérateur Θ est liée à l'inversion de l'opérateur Γ. Ce problème inverse étant mal posé, de façon analogue à ce qui a été présenté dans le chapitre 2, nous considérons un entier k_n permettant de tronquer la somme dans l'inversion de l'opérateur $\widehat{\Gamma}_n$. Rappelons que $(\widehat{\lambda}_j)_{j\geq 1}$ et $(\widehat{v}_j)_{j\geq 1}$ désignent respectivement la suite des valeurs propres et la suite des fonctions propres de l'opérateur de covariance empirique $\widehat{\Gamma}_n$, et $\widehat{\Pi}_{k_n}$ désigne l'opérateur de projection sur le sous-espace $\text{span}(\widehat{v}_1, \dots, \widehat{v}_{k_n})$. Ainsi, l'estimation de l'opérateur Θ est donnée par

$$\widehat{\Theta} = \widehat{\Pi}_{k_n}\widehat{\Delta}_n\left(\widehat{\Pi}_{k_n}\widehat{\Gamma}_n\right)^{-1}, \tag{3.8}$$

et l'estimation équivalente de la fonction θ est donnée par

$$\widehat{\theta}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i, \widehat{v}_j\rangle}{\widehat{\lambda}_j}Y_i(t)\widehat{v}_j(s), \tag{3.9}$$

comme cela a été considéré dans le chapitre 2, dans le cas où la variable réponse est réelle.

À partir de l'estimation de Θ, nous pouvons définir la prévision de Y_{n+1}, étant donnée une nouvelle entrée X_{n+1}, par

$$\widehat{Y}_{n+1}(t) = \widehat{\Theta} X_{n+1}(t) = \int_0^1 \widehat{\theta}(s,t) X_{n+1}(s) \mathrm{d}s. \tag{3.10}$$

3.3 Résultats asymptotiques

Les preuves des résultats qui suivent sont données dans Crambes et Mas (2013).

3.3.1 Hypothèses

Les hypothèses dont nous avons besoin sont de trois types : régularité de l'opérateur Θ, hypothèses de moments sur X et hypothèses de régularité sur X par le biais d'une expression de la décroissance vers zéro des valeurs propres de l'opérateur de covariance Γ.

Hypothèse sur Θ

Nous supposons que Θ est un opérateur de Hilbert-Schmidt : pour toute base $(e_j)_{j\geq 1}$ de H, on a

$$\sum_{j,\ell=1}^{+\infty} \langle \Theta e_\ell, e_j \rangle^2 < +\infty. \tag{3.11}$$

Cette hypothèse est équivalente au fait de supposer que θ est doublement intégrable sur H.

Hypothèse de moment sur X

Rappelons d'abord le développement de Karhunen-Loève de X, qui n'est autre que la décomposition de X dans la base $(v_j)_{j\geq 1}$ des fonctions propres de l'opérateur de covariance Γ

$$X = \sum_{j=1}^{+\infty} \sqrt{\lambda_j} \xi_j v_j \quad p.s.,$$

où les ξ_j's sont des variables aléatoires réelles centrées et non corrélées, de variance unitaire. Nous supposons que pour $j, \ell \geq 1$ il existe une constante b telle que

$$\mathbb{E}\left(|\xi_j|^\ell\right) \leq \frac{\ell!}{2} b^{\ell-2} \mathbb{E}\left(|\xi_j|^2\right). \tag{3.12}$$

Cette hypothèse fait écho à l'hypothèse (2.19) p.49 dans Bosq (2000). Le but de cette hypothèse est de pouvoir appliquer des inégalités exponentielles de type Bernstein. De plus, l'hypothèse (3.12) a comme conséquence

$$\mathbb{E}\langle X, v_j\rangle^4 \leq C\left(\mathbb{E}\langle X, v_j\rangle^2\right)^2. \tag{3.13}$$

Cette condition relativement classique, indique que la suite des moments d'ordre 4 des marginales de X tend vers zéro suffisamment vite. L'hypothèse (3.12) est vérifiée par exemple lorsque X est un processus gaussien, ou encore lorsque X est borné p.s.

Hypothèse sur les valeurs propres de Γ

Soit la fonction $\lambda : \mathbb{R}^+ \to \mathbb{R}^+$ définie par $\lambda(j) = \lambda_j$ pour tout $j \geq 1$ et interpolée continûment entre j et $j+1$. Nous supposons que

$$x \to \lambda(x) \text{ est convexe.} \tag{3.14}$$

Cette condition est faible et couvre un ensemble très large de décroissance de valeurs propres : décroissance polynômiale $\lambda_j = C_\alpha j^{-1-\alpha}$ avec $\alpha > 0$ ou décroissance exponentielle $\lambda_j = D_\alpha j^\beta \exp(-\alpha j)$. De telles décroissances permettent de considérer des processus allant du très irrégulier (même plus irréguliers que le mouvement Brownien pour lequel $\lambda_j = C_1 j^{-2}$) au très lisse.

3.3.2 Erreur de prédiction en moyenne quadratique

Nous commençons par donner une borne supérieure, puis le risque asymptotique exact du prédicteur. Soit $\Gamma_\varepsilon = \mathbb{E}(\langle\varepsilon, .\rangle\varepsilon)$ l'opérateur de covariance de l'erreur de modèle et $\sigma_\varepsilon^2 = \operatorname{tr}(\Gamma_\varepsilon)$.

Theorème 3.1. *L'erreur moyenne quadratique de prédiction a le développement asymptotique :*

$$\mathbb{E}\left\|\widehat{\Theta}_n X_{n+1} - \Theta X_{n+1}\right\|^2 = \sigma_\varepsilon^2 \frac{k}{n} + \sum_{j=k+1}^{+\infty} \lambda_j \|\Theta v_j\|^2 + A_n + B_n, \tag{3.15}$$

où $A_n \leq C_A \frac{k^2\lambda_k}{n} \|\Theta\|_{\mathcal{L}_2}$ *et* $B_n \leq C_B \frac{k^2(\log k)}{n^2}$ *avec* C_A *et* C_B *constantes indépendantes de* k, n *et* Θ.

Les deux premiers termes déterminent la vitesse de convergence : la variance apparaît à travers le terme $\sigma_\varepsilon^2 \frac{k}{n}$ et le biais au carré à travers le terme $\sum_{j=k+1}^{+\infty} \lambda_j \|\Theta v_j\|^2$. Le terme A_n provient de la décomposition du terme de biais au carré et B_n est un terme résiduel de la variance. Les deux sont négligeables devant les deux premiers termes du développement asymptotique. Une propriété intéressante provient du Théorème 3.1. En écrivant $\lambda_j \|\Theta v_j\|^2 = \left\|\Gamma^{1/2}\Theta v_j\right\|^2$, nous pouvons voir que les seules conditions de régularité requises proviennent de la décomposition spectrale de l'opérateur $\Gamma^{1/2}\Theta$ et non de Γ et de Θ séparément.

3.3.3 Optimalité

Avant de présenter des résultats d'optimalité, nous introduisons la classe de paramètres Θ sur laquelle l'optimalité est obtenue.

Définition 3.1. *Soit $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ une fonction décroissante de classe C^1 telle que $\sum_{j=1}^{+\infty} \varphi(j) = 1$. Nous notons $\mathcal{L}_2(\varphi, L)$ la classe des opérateurs linéaires de H dans H définie par*

$$\mathcal{L}_2(\varphi, L) = \left\{ T \in \mathcal{L}_2, \|T\|_{\mathcal{L}_2} \leq L : \|Tv_j\| \leq L\sqrt{\varphi(j)} \right\}.$$

L'ensemble $\mathcal{L}_2(\varphi, L)$ est entièrement déterminé par la constante L et la fonction φ. Nous pouvons alors utiliser le résultat du Théorème 3.1 afin de choisir k de telle sorte que les termes principaux soient du même ordre de grandeur. Par conséquent, nous obtenons une borne uniforme sur l'espace $\mathcal{L}_2(\varphi, L)$ de la vitesse de convergence pour l'erreur de prédiction.

Theorème 3.2. *Soit $L = \left\|\Theta\Gamma^{1/2}\right\|_{\mathcal{L}_2}$, $\varphi(j) = \lambda_j \|\Theta(e_j)\|^2 / L^2$. On note k_n^* la partie entière de l'unique solution en x de l'équation*

$$\frac{1}{x}\int_x^{+\infty} \varphi(x)\mathrm{d}x = \frac{1}{n}\frac{\sigma_\varepsilon^2}{L^2}. \tag{3.16}$$

Nous avons alors

$$\limsup_{n\to+\infty} \frac{n}{k_n^*} \sup_{\Theta\Gamma^{1/2}\in\mathcal{L}_2(L,\varphi)} \mathbb{E}\left\|\widehat{\Theta}_n(X_{n+1}) - \Theta(X_{n+1})\right\|^2 = 2\sigma_\varepsilon^2.$$

Comme cela a déjà été souligné dans le chapitre précédent dans le cadre des données manquantes sur la variable réponse réelle, l'équation (3.16) a une unique solution. L'entier k_n^* peut être vu comme une dimension optimale, minimisant l'erreur de prédiction. Également, il est possible d'expliciter des vitesses de convergence lorsque la décroissance de la fonction φ est particulière (polynômiale ou exponentielle). Nous donnons ces résultats dans le corollaire suivant.

Corollaire 3.3. *Nous considérons, comme dans l'exemple 2.1 du chapître 1, les fonctions $\varphi_{pol}(j) = C_\alpha j^{-(2+\alpha)}$ et $\varphi_{exp}(j) = D_\alpha \exp(-\alpha j)$ où C_α et D_α sont des constantes positives et $\alpha > 0$. Alors, la solution de l'équation (3.16) vérifie*

$$\begin{cases} k_{n,pol}^\star \underset{n\to+\infty}{\sim} \left(\frac{C_\alpha L^2}{(1+\alpha)\sigma_\varepsilon^2}\right)^{1/(2+\alpha)} n^{1/(2+\alpha)}, & si\ \varphi = \varphi_{pol}, \\ k_{n,exp}^\star \underset{n\to+\infty}{\lesssim} \frac{\log n}{\alpha}, & si\ \varphi = \varphi_{exp}. \end{cases}$$

Pour $\varphi = \varphi_{pol}$, le résultat du Théorème 3.2 devient

$$\sup_{\Theta\Gamma^{1/2}\in\mathcal{L}_2(L,\varphi)} \mathbb{E}\left\|\widehat{\Theta}_n\left(X_{n+1}\right) - \Theta\left(X_{n+1}\right)\right\|^2 \underset{n\to+\infty}{\sim} 2\left(\sigma_\varepsilon^2\right)^{(1+\alpha)/(2+\alpha)} \left(\frac{C_\alpha L^2}{1+\alpha}\right)^{1/(2+\alpha)} \frac{n^{1/(2+\alpha)}}{n-m_n},$$

et pour $\varphi = \varphi_{exp}$, le résultat du Théorème 3.2 devient

$$\sup_{\Theta\Gamma^{1/2}\in\mathcal{L}_2(L,\varphi)} \mathbb{E}\left\|\widehat{\Theta}_n\left(X_{n+1}\right) - \Theta\left(X_{n+1}\right)\right\|^2 \underset{n\to+\infty}{\lesssim} \frac{2\sigma_\varepsilon^2 \log n}{\alpha(n-m_n)}.$$

Comme cela a été remarqué dans le chapitre 2, les vitesses obtenues correspondent à des vitesses optimales usuelles dans ce contexte. Pour aller au bout de l'étude de l'optimalité de ces vitesses, nous donnons le résultat suivant.

Theorème 3.4. *La borne suivante prouve que notre estimateur est optimal dans un sens minimax :*

$$\inf_{\widehat{\Theta}_n} \sup_{\Theta\Gamma^{1/2}\in\mathcal{L}_2(\varphi,L)} \mathbb{E}\left\|\widehat{\Theta}_n\left(X_{n+1}\right) - \Theta\left(X_{n+1}\right)\right\|^2 \sim \frac{k_n^*}{n}.$$

3.3.4 Convergence en loi

Dans ce paragraphe, nous présentons des résultats de convergence faible. Nous commençons avec un résultat négatif qui montre que le problème de convergence faible n'a pas de solution pour des topologies trop fortes.

Theorème 3.5. *L'estimateur $\widehat{\Theta}_n$ ne peut pas converger en distribution pour la norme de Hilbert-Schmidt.*

Nous donnons à présent un résultat plus positif, qui améliore les résultats de Cardot et al. (2007). La convergence faible est notée $\xrightarrow{w}$. Deux résultats sont prouvés. Le premier donne la convergence faible du prédicteur avec un terme de biais. Le second élimine ce biais sous une condition plus forte sur k_n.

Theorème 3.6. *Si la condition $(k\log k)^2/n \to 0$ est vérifiée, alors*

$$\sqrt{\frac{n}{k}}\left[\widehat{\theta}_n X_{n+1} - \Theta\Pi_k X_{n+1}\right] \xrightarrow{w} \mathcal{G}_\varepsilon$$

où $\mathcal{G}_\varepsilon$ est un élément de H gaussien centré d'opérateur de covariance Γ_ε. De plus, si nous notons $\gamma_k = \sup_{j\geq k}\left\{j\log j\left\|\Theta v_j\right\|\sqrt{\lambda_j}\right\}$ et si nous choisissons k tel que $n \leq (k\log k)^2/\gamma_k$ (ce qui signifie que $(k\log k)^2/n$ ne décroît pas trop vite vers zéro), le terme de biais disparaît et on obtient

$$\sqrt{\frac{n}{k}}\left[\widehat{\Theta}_n X_{n+1} - \Theta X_{n+1}\right] \xrightarrow{w} \mathcal{G}_\varepsilon.$$

Une application immédiate du Théorème 3.6 est l'obtention d'intervalles de prédiction asymptotiques. La notation Y_{n+1}^* correspond à $\Theta X_{n+1} = \mathbb{E}\left(Y_{n+1}|X_{n+1}\right)$.

Corollaire 3.7. *Soit m une fonction fixée dans $H = L^2([0,1])$. Nous présentons un intervalle de prédiction asymptotique pour $\int Y_{n+1}^*(t)\, m(t)\, dt$ au niveau $1-\alpha$:*

$$\mathbb{P}\left(\int_0^1 Y_{n+1}^*(t)\, m(t)\, dt \in \left[\int_0^1 \widehat{Y}_{n+1}(t)\, m(t)\, dt \pm \sqrt{\frac{k}{n}}\sigma_m q_{1-\alpha/2}\right]\right) = 1-\alpha,$$

où $\sigma_m^2 = \langle m, \Gamma_\varepsilon m\rangle = \int\int \Gamma_\varepsilon(s,t)\, m(t)\, m(s)\, dtds$ et $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite $\mathcal{N}(0,1)$.

Pour obtenir un intervalle de prédiction pour $Y_{n+1}^*(t_0)$ (avec t_0 fixé dans $[0,1]$), nous devons nous assurer que $f \in H \longmapsto f(t_0)$ est continue pour la norme $\|.\|$. Cette fonctionnelle est toujours continue dans l'espace $\left(C([0,1]), |.|_\infty\right)$ mais pas dans l'espace $L^2([0,1])$. Un changement d'espace H amène au résultat.

Corollaire 3.8. *Considérons*

$$H = W_0^{2,1}([0,1]) = \left\{f \in L^2([0,1]) : f(0) = 0, f' \in L^2([0,1])\right\},$$

muni de son produit scalaire $\langle u, v\rangle = \int_0^1 u'v'$, l'évaluation $f \in H \longmapsto f(t_0)$ est continue pour la norme de H et on a

$$\mathbb{P}\left(Y_{n+1}^*(t_0) \in \left[\widehat{Y}_{n+1}(t_0) \pm \sqrt{\frac{k}{n}}\sigma_{t_0} q_{1-\alpha/2}\right]\right) = 1-\alpha,$$

où $\sigma_{t_0}^2 = \Gamma_\varepsilon(t_0, t_0)$.

3.4 Estimation de l'opérateur de covariance de l'erreur

Dans cette section, nous nous intéressons à l'estimation de l'opérateur de covariance de l'erreur. Le résultat du Théorème 3.1 fait apparaître l'intérêt de cette estimation à travers le terme σ_ε^2. Nous nous intéressons à deux types d'estimateurs, que nous comparons ensuite. Les preuves des résultats sont données dans Crambes et al. (2016).

3.4.1 Estimateur Plug-in

L'estimateur plug-in de Γ_ε est défini par

$$\widehat{\Gamma}_\varepsilon = \frac{1}{n-k_n}\sum_{i=1}^n \langle Y_i - \widehat{\Theta}_n X_i, .\rangle (Y_i - \widehat{\Theta}_n X_i) = \frac{1}{n-k_n}\sum_{i=1}^n \langle \widehat{\varepsilon}_i, .\rangle \widehat{\varepsilon}_i. \tag{3.17}$$

Cet estimateur est biaisé, à n fixé, d'après le résultat suivant.

Theorème 3.9. *Nous avons le développement*

$$\mathbb{E}\left[\widehat{\Gamma}_{\varepsilon}\right] = \Gamma_{\varepsilon} + \left(\frac{n}{n-k_n}\right)\Theta\mathbb{E}\left(\sum_{i=k_n+1}^{n}\widehat{\lambda}_i\langle\widehat{v}_i,.\rangle\widehat{v}_i\right)\Theta'. \tag{3.18}$$

Comme $\widehat{\Gamma}_n = \sum_{i=1}^{n}\widehat{\lambda}_i\langle\widehat{v}_i,.\rangle\widehat{v}_i$ et $\widehat{\Pi}_{(k_n+1):n} := \sum_{i=k_n+1}^{n}\langle\widehat{v}_i,.\rangle\widehat{v}_i$, nous pouvons déduire le résultat suivant.

Corollaire 3.10. *Nous avons le développement*

$$\mathbb{E}\left[\widehat{\Gamma}_{\varepsilon}\right] = \Gamma_{\varepsilon} + \left(\frac{n}{n-k_n}\right)\Theta\mathbb{E}\left(\widehat{\Pi}_{(k_n+1):n}\widehat{\Gamma}_n\right)\Theta'. \tag{3.19}$$

Sous des conditions simples, nous prouvons que l'estimateur plug-in (3.17) de Γ_{ε} est asymptotiquement non biaisé. Les hypothèses que nous considérons sont les suivantes.

$(A.1)$ L'operateur Θ est un opérateur nucléaire, soit $\|\Theta\|_{\mathcal{N}} < +\infty$.

$(A.2)$ La variable X satisfait $\mathbb{E}\|X\|^4 < +\infty$.

$(A.3)$ Nous avons presque sûrement $\widehat{\lambda}_1 > \widehat{\lambda}_2 > \ldots > \widehat{\lambda}_{k_n} > 0$.

$(A.4)$ Nous avons $\lambda_1 > \lambda_2 > \ldots > 0$.

Nous pouvons maintenant énoncer le résultat.

Theorème 3.11. *Sous les hypothèses* $(A.1)$-$(A.4)$, *si* $(k_n)_{n\geq 1}$ *vérifie* $\lim_{n\to+\infty} k_n = +\infty$ *et* $\lim_{n\to+\infty} k_n/n = 0$, *nous avons*

$$\lim_{n\to+\infty}\left\|\mathbb{E}\left(\widehat{\Gamma}_{\varepsilon}\right) - \Gamma_{\varepsilon}\right\|_{\mathcal{N}} = 0. \tag{3.20}$$

Ce résultat a la conséquence immédiate suivante.

Corollaire 3.12. *Sous les hypothèses du Théorème 3.11, nous avons*

$$\lim_{n\to+\infty}\mathbb{E}\left[tr\left(\widehat{\Gamma}_{\varepsilon}\right)\right] = tr\left(\Gamma_{\varepsilon}\right). \tag{3.21}$$

3.4.2 Correction du biais

Sans perte de généralté, nous allons considérer un nombre d'observations n multiple de 3. Dans la relation (3.19), le biais de l'estimateur est lié à $\Theta\mathbb{E}\left(\widehat{\Pi}_{(k_n+1):n}\widehat{\Gamma}_n\right)\Theta'$. Une autre façon d'estimer Γ_{ε} est d'enlever une estimation du biais à l'estiamteur plug-in $\widehat{\Gamma}_{\varepsilon}$. Pour cela, nous partageons l'échantillon de taille n en trois sous-échantillons de taille $m = n/3$. Par conséquent, nous définissons

$$\breve{B}_n = \widehat{\Theta}^{[2]}_{2k_m}\left(\widehat{\Pi}^{[1]}_{(k_m+1):m}\widehat{\Gamma}^{[1]}_{m}\right)\left(\widehat{\Theta}^{[3]}_{2k_m}\right)', \tag{3.22}$$

où les quantités avec des exposants [1], [2] et [3] sont respectivement estimées avec la première, deuxième et troisième partie de l'échantillon. Nous utilisons $2k_m$ valeurs propres (avec $k_m \leq n/2$) dans l'estimation de Θ avec le deuxième et le troisième sous-échantillon pour éviter l'orthogonalité entre $\widehat{\Theta}^{[2]}_{2k_m}$, $\widehat{\theta}^{[3]}_{2k_m}$ et $\widehat{\Pi}^{[1]}_{(k_m+1):m}\widehat{\Gamma}^{[1]}_{m}$. Nous pouvons à présent définir l'estimateur de Γ_ε :

$$\breve{\Gamma}_\varepsilon = \widehat{\Gamma}^{[1]}_\varepsilon - \frac{m}{m-k_m}\breve{B}_n, \tag{3.23}$$

où $\widehat{\Gamma}^{[1]}_\varepsilon$ est l'estimateur plug-in de Γ_ε basé sur le premier sous-échantillon. Nous pouvons à présent énoncer le résultat suivant.

Theorème 3.13. *Sous les hypothèses du Théorème 3.11, nous avons*

$$\lim_{n\to+\infty} \left\| \mathbb{E}\left(\breve{\Gamma}_\varepsilon\right) - \Gamma_\varepsilon \right\|_{\mathcal{N}} = 0. \tag{3.24}$$

Ce résultat a la conséquence immédiate suivante.

Corollaire 3.14. *Sous les hypothèses du Théorème 3.11, nous avons*

$$\lim_{n\to+\infty} \mathbb{E}\left[tr\left(\breve{\Gamma}_\varepsilon\right)\right] = tr\left(\Gamma_\varepsilon\right). \tag{3.25}$$

3.4.3 Commentaires sur les deux estimateurs

Le fait de soustraire un estimateur du biais de l'estimateur plug-in $\widehat{\Gamma}_\varepsilon$ ne permet pas d'obtenir un estimateur non biaisé de Γ_ε (du moins, il ne semble pas immédiat de le prouver). La situation est complètement différente de celle des modèles de régression multiple multivariée (voir Johnson et Wichern (2007)), où un estimateur non biaisé de la matrice de variance-covariance du bruit est facilement construit.

Les deux estimateurs $\widehat{\Gamma}_\varepsilon$ et $\breve{\Gamma}_\varepsilon$ sont consistants. Cependant, il ne semble pas possible dans l'immédiat de prouver que soustraire le biais améliore l'estimation de Γ_ε, ni de sa trace. Une comparaison pratique entre les deux estimateurs est menée dans le paragraphe suivant.

3.4.4 Simulations

La variable X est simulée comme un mouvement Brownien standard sur $[0,1]$, observé aux temps $[\frac{1}{1000}, \frac{2}{1000}, \ldots, \frac{1000}{1000}]$. Nous avons considéré des tailles d'échantillon $n = 300$ et $n = 1\,500$. Nous avons simulé le bruit ε comme un mouvement Brownien standard multiplié par 0.1 pour avoir un rapport signal sur bruit de 10. La trace de l'opérateur de covariance de ε est $tr(\Gamma_\varepsilon) = 0.005$.

Nous avons également considéré deux types d'opérateurs Θ. Le premier est $\Theta = \Pi_{20} = \sum_{j=1}^{20}\langle v_j, .\rangle v_j$ où les v_j sont les fonctions propres de l'opérateur de covariance de

X (donc connues pour un mouvement Brownien standard : $v_j(t) = \sqrt{2}\sin((j-1/2)\pi t)$). Le deuxième opérateur Θ considéré est l'opérateur intégral défini par $\Theta X = \int_0^1 (t^2 + s^2)X(s)ds$.

Dans la suite, nous nous intéressons aux deux estimateurs définis précédemment, l'estimateur plug-in défini dans (3.17) et à l'estimateur avec correction du biais défini dans (3.22) et (3.23). Nous avons choisi de considérer un troisième estimateur

$$\widehat{\breve{\Gamma}}_\varepsilon = \widehat{\Gamma}_\varepsilon - \left(\frac{n}{n-k_n}\right)\left[\widehat{\Theta}_n(\widehat{\Pi}_{(k_n+1):n}\widehat{\Gamma}_n)(\widehat{\Theta}_n)'\right].$$

Ce troisième estimateur utilise tout l'échantillon pour corriger le biais, il n'est donc pas possible d'avoir un résultat de consistance sur cet estimateur (car l'indépendance entre les termes intervenant dans la construction de l'estimateur disparaît). Cependant, nous souhaitons voir son comportement en pratique.

Nous présentons dans la table 3.1 (simulation 1) et dans la table 3.2 (simulation 2) les valeurs moyennes et les écarts-types de la trace obtenus pour les trois estimateurs sur $N = 100$ simulations. Nous donnons également les valeurs des critères de validation croisée (CV) et de validation croisée généralisée (GCV).

n	k	$CV(k)$	$GCV(k)$	$tr(\widehat{\Gamma}_\varepsilon)$	$tr(\breve{\Gamma}_\varepsilon)$	$tr(\widehat{\breve{\Gamma}}_\varepsilon)$
n=300	16	6.67 (3.5)	6.67 (3.5)	6.32 (3.3)	5.29 (7.1)	4.79 (3.3)
	18	6.04 (3.5)	6.04 (3.5)	5.67 (3.3)	5.13 (7.2)	4.74 (3.4)
	20	5.66 (3.7)	5.66 (3.7)	5.28 (3.4)	5.06 (7.1)	4.7 (3.4)
	22	**5.5698 (3.7)**	5.57 (3.6)	5.16 (3.4)	5.04 (7.1)	4.67 (3.4)
	24	5.57 (3.7)	**5.568 (3.7)**	5.12 (3.4)	5.02 (7.1)	4.63 (3.4)
	26	5.59 (3.8)	5.59 (3.8)	5.11 (3.4)	5.02 (7.2)	4.58 (3.4)
n=1500	18	5.67 (1.7)	5.67 (1.7)	5.6 (1.7)	5.03 (2.5)	4.97 (1.7)
	20	5.15 (1.7)	5.15 (1.7)	5.08 (1.7)	5.01 (2.5)	4.96 (1.7)
	22	5.12 (1.7)	5.12 (1.7)	5.04 (1.7)	5.01 (2.6)	4.95 (1.7)
	24	**5.11 (1.7)**	**5.11 (1.7)**	5.04 (1.7)	5.01 (2.6)	4.95 (1.7)
	26	5.12 (1.7)	5.12 (1.7)	5.03 (1.7)	5 (2.6)	4.94 (1.7)
	28	5.13 (1.7)	5.13 (1.7)	5.03 (1.7)	5 (2.6)	4.93 (1.7)

Table 3.1 Critères CV et GCV pour différentes valeurs de k, valeurs moyennes et écarts-types des estimateurs de $Tr(\Gamma_\varepsilon)$ (simulation 1 avec $n = 300$ and $n = 1500$). Les valeurs sont données multipliées par 10^3 (les écarts-types entre parenthèses sont donnés multipliés par 10^4).

Dans la première simulation, la vraie valeur de k est connue ($k = 20$), les valeurs choisies par CV et GCV sont $k = 22$ ou $k = 24$. Pour ces valeurs de k, le meilleur estimateur est $tr(\breve{\Gamma}_\varepsilon)$ pour $n = 300$ et $n = 1\,500$. La surestimation de $tr(\widehat{\Gamma}_\varepsilon)$ semble être bien corrigée par $tr(\breve{\Gamma}_\varepsilon)$, même si l'utilité de la correction du biais ne peut pas

être prouvée théoriquement pour l'instant. Sur cette simulation, l'estimateur $tr(\widehat{\breve{\Gamma}}_\varepsilon)$ ne semble pas meilleur que les autres, en particulier pour une taille d'échantillon plus petite.

n	k	$CV(k)$	$GCV(k)$	$tr(\widehat{\Gamma}_\varepsilon)$	$tr(\breve{\Gamma}_\varepsilon)$	$tr(\widehat{\breve{\Gamma}}_\varepsilon)$
n=300	2	5.37 (3.6)	5.37 (3.6)	5.34 (3.6)	5.03 (6.4)	5.07 (3.2)
	4	**5.17 (3.3)**	**5.17 (3.3)**	5.11 (3.2)	5.02 (6.4)	5 (3.1)
	6	5.18 (3.2)	5.18 (3.2)	5.08 (3.2)	5 (6.5)	4.96 (3.2)
	8	5.21 (3.2)	5.21 (3.2)	5.07 (3.2)	5 (6.4)	4.93 (3.2)
	10	5.25 (3.3)	5.25 (3.3)	5.07 (3.2)	5 (6.6)	4.89 (3.2)
n=1500	2	5.28 (1.7)	5.28 (1.7)	5.28 (1.7)	5.04 (2.8)	5.05 (1.7)
	4	5.07 (1.7)	5.07 (1.7)	5.05 (1.7)	5.01 (2.6)	5.02 (1.7)
	6	**5.05 (1.7)**	**5.05 (1.7)**	5.03 (1.7)	5 (2.6)	5.01 (1.7)
	8	5.06 (1.7)	5.06 (1.7)	5.03 (1.7)	5 (2.5)	5 (1.7)
	10	5.06 (1.7)	5.06 (1.7)	5.03 (1.7)	5 (2.5)	4.99 (1.7)

Table 3.2 Critères CV et GCV pour différentes valeurs de k, valeurs moyennes et écarts-types des estimateurs de $Tr(\Gamma_\varepsilon)$ (simulation 2 avec $n = 300$ and $n = 1500$). Les valeurs sont données multipliées par 10^3 (les écarts-types entre parenthèses sont donnés multipliés par 10^4).

Dans la deuxième simulation, la vraie valeur de k est inconnue, et la valeur choisie par les critères CV et GCV est $k = 4$ (pour $n = 300$) ou $k = 6$ (pour $n = 1\,500$). L'estimateur $tr(\widehat{\breve{\Gamma}}_\varepsilon)$ est légèrement meilleur que les deux autres estimateurs pour $n = 300$. Pour $n = 1500$, $tr(\breve{\Gamma}_\varepsilon)$ est légèrement meilleur.

Sur les deux simulations, $tr(\widehat{\Gamma}_\varepsilon)$ et $tr(\breve{\Gamma}_\varepsilon)$ montrent une bonne précision d'estimation. D'un point de vue pratique, $tr(\widehat{\Gamma}_\varepsilon)$ peut être préféré dans la mesure où il est très facile à implémenter. La correction du biais de $tr(\breve{\Gamma}_\varepsilon)$ donnera une estimation plus précise dans les cas où cela est nécessaire.

3.5 Perspectives

La principale perspective de ce travail a été évoquée à la fin du chapitre 2 : il s'agit d'étudier le modèle linéaire fonctionnel avec sortie fonctionnelle lorsque des données manquantes apparaissent.

Bibliographie

A. Aguilera, F. Ocaña et M. Valderrama. : Estimation of functional regression models for functional responses by wavelet approximat. In *International Workshop on Functional and Operatorial Statistics 2008 Proceedings, Functional and operatorial statistics, Dabo-Niang and Ferraty (Eds)*. Physica-Verlag, Springer, 2008.

J. Antoch, L. Prchal, M. De Rosa et P. Sarda. Functional linear regression with functional response : application to prediction of electricity consumpt. In *International Workshop on Functional and Operatorial Statistics 2008 Proceedings, Functional and operatorial statistics, Dabo-Niang and Ferraty (Eds)*. Physica-Verlag, Springer, 2008.

D. Benatia, M. Carrasco et J.-P. Florens. Functional linear regression with functional response. *Journal of Econometrics*, 201 :269–291, 2017.

D. Bosq. *Linear Processues in Function Spaces : Theory and Applications (Lecture notes in Statistics)*. Springer-verlaog New York, 2000.

H. Cardot, A. Mas et P. Sarda. Clt in functional linear regression models. *Probability Theory and Related Fields*, 138 :325–361, 2007.

J.M. Chiou, H.G. Müller et J.L. Wang. Functional response models. *Statistica Sinica*, 14 :675–693, 2004.

C. Crambes et A. Mas. Asymptotics of prediction in functional linear regression with functional outputs. *Bernoulli*, 19 :2627–2651, 2013.

C. Crambes, N. Hilgert et T. Manrique. Estimation of the noise covariance operator in functional linear regression with functional outputs. *Statistics and Probability Letters*, 113 :7–15, 2016.

A. Cuevas, M. Febrero et R. Fraiman. Linear functional regression : the case of fixed design and functional response. *Canadian Journal of Statistics*, 30(2) :285–300, 2002.

M. Imaizumi et K. Kato. Pca-based estimation for functional linear regression with functional responses. *Journal of Multivariate Analysis*, 163 :15–36, 2018.

Richard Arnold Johnson et Dean W Wichern. *Applied multivariate statistical analysis*. Prentice Hall Englewood Cliffs, NJ, sixth edition, 2007.

J.O. Ramsay et B.W. Silverman. *Functional Data Analysis (Second edition)*. Springer-Verlag, New York, 2005.

F. Yao, H.G. Müller et J.L. Wang. Functional linear regression analysis for longitudinal data. *Annals of Statistics*, 33 :2873–2903, 2005.

Chapitre 4

Modèle fonctionnel de convolution

Contents

Les résultats présentés dans ce chapitre sont tirés de Manrique et al. (2018) et Manrique et al. (2019).

4.1 Introduction

Dans ce chapitre, nous nous intéressons au modèle de convolution

$$Y(t) = \int_0^t \theta(s)\, X(t-s)ds + \varepsilon(t), \tag{4.1}$$

où $t \geq 0$, θ est la fonction inconnue à estimer, X, Y sont des fonctions aléatoires et ε est un bruit aléatoire fonctionnel. Toutes ces fonctions sont supposées nulles pour $t < 0$. Ce modèle est un sous-modèle du modèle fonctionnel historique

$$Y(t) = \int_0^t \mathcal{H}(s,t)\, X(s)ds + \varepsilon(t), \tag{4.2}$$

lui même étant un sous-modèle du modèle complètement fonctionnel

$$Y(t) = \int \theta(s,t)\, X(s)ds + \varepsilon(t), \tag{4.3}$$

déjà introduit dans le chapitre 3. Le modèle fonctionnel historique a été introduit par Malfait et Ramsay (2003) afin de pouvoir considérer des situations spécifiques (notamment où les variables sont temporelles) pour lesquelles la variable Y en l'instant t est expliquée par le passé de la variable X entre 0 et t. Dans notre cas, nous supposons une forme plus simple au noyau $\mathcal{H}(s,t)$ sous la forme $\theta(t-s)$. Nous supposons donc que la fonction à estimer n'est plus que la fonction d'une seule variable.

Certains travaux sont liés à l'étude du modèle (4.1) ou à ses dérivés. Par exemple, Asencio et al. (2014) étudient un problème similaire avec plusieurs variables explicatives fonctionnelles. L'estimation de θ se fait par projection sur des sous-espaces de fonctions splines. Par ailleurs, Malfait et Ramsay (2003) considèrent le cas où la fonction θ dans le modèle (4.1) est une fonction de deux variables (soit $\theta(s,t)$). Ainsi, l'intégrale devient un opérateur à noyau et ils utilisent la méthode des éléments finis pour estimer θ sur un certain domaine. Ces deux articles sont les plus pertinents dans la littérature concernant l'estimation de θ dans le modèle (4.1) à partir d'un échantillon $(X_i, Y_i)_{i=1,\ldots,n}$. Nous avons proposé une approche différente en utilisant une représentation équivalente de ce modèle dans le domaine des fréquences. Nous adoptons une stratégie d'estimation où l'idée principale est d'utiliser la transformée de Fourier continue du modèle (4.1). Nous pouvons ainsi tranformer le modèle de départ (domaine temporel) en un modèle plus simple, connu sous le nom de modèle fonctionnel de concurrence (domaine fréquentiel). Ce modèle est notamment présenté sous une forme plus générale dans Ramsay et Silverman (2005).

Une fois dans le domaine fréquentiel, nous estimons la fonction inconnue du modèle à l'aide d'un estimateur de type ridge. Nous obtenons des résultats de consistence de cet estimateur, qu'il est ensuite possible de transposer à l'estimation de la fonction θ en revenant dans le domaine temporel par transformée de Fourier inverse.

4.2 Modèles fonctionnels de convolution et de concurrence

4.2.1 Notations

Commençons par donner quelques notations utiles pour toute la suite. Nous définissons $L^1(\mathbb{R},\mathbb{C})$ l'espace des fonctions intégrables à valeurs complexes, muni de la norme L^1 définie par $\|f\|_{L^1} := \left[\int_{\mathbb{R}} |f(x)|dx\right]$ où $|\cdot|$ désigne le module complexe. Nous définissons $L^1(\mathbb{R},\mathbb{R})$ l'espace des fonctions intégrables à valeurs réelles. De façon analogue, $L^2(\mathbb{R},\mathbb{C})$ est l'espace des fonctions de carré intégrable à valeurs complexes, muni de la norme L^2 définie par $\|f\|_{L^2} := \left[\int_{\mathbb{R}} |f(x)|^2 dx\right]^{1/2}$, et $L^2(\mathbb{R},\mathbb{R})$ est l'espace des fonctions de carré intégrable à valeurs réelles. Étant donné un sous-ensemble $K \subset \mathbb{R}$, $\|f\|_{L^2(K)} := \left[\int_K |f(x)|^2 dx\right]^{1/2}$.

Soit $C_0(\mathbb{R},\mathbb{C}) = C_0$ l'espace des fonctions f continues à valeurs complexes qui tendent vers zéro à l'infini, c'est-à-dire que pour tout $\zeta > 0$, il existe $R > 0$ tel que pour tout $|t| > R$, on a $|f(t)| < \zeta$. Nous utilisons la norme supremum définie par $\|f\|_{C_0} := \sup_{x\in\mathbb{R}} |f(x)|$. Étant donné un sous-ensemble $K \subset \mathbb{R}$, cette norme devient $\|f\|_{C_0(K)} := \sup_{x\in K} |f(x)|$.

La transformée de Fourier continue est notée $\mathcal{F}$ et son inverse $\mathcal{F}^{-1}$, définies respectivement par

$$\mathcal{F}(f) : \xi \longmapsto \mathcal{F}(f)(\xi) = \int_{-\infty}^{+\infty} f(x) e^{-i\xi x} dx,$$

et

$$\mathcal{F}^{-1}(g) : x \longmapsto \mathcal{F}^{-1}(g)(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} g(\xi) e^{ix\xi} d\xi.$$

Enfin, tout au long de ce chapitre, le support d'une fonction continue $f : \mathbb{R} \to \mathbb{C}$ est l'ensemble $supp(f) := \{t \in \mathbb{R} : |f(t)| \neq 0\}$. Nous définissons également la frontière d'un ensemble S par $\partial(S) := \overline{S} \setminus int(S)$, où $\overline{S}$ est la fermeture de S et $int(S)$ est son intérieur.

4.2.2 Hypothèses générales

Nous donnons ici les hypothèses sur le modèle.

$(HA1_{FCVM})$ Les variables X et ε sont indépendantes, à valeurs dans $L^1(\mathbb{R},\mathbb{R}) \cap L^2(\mathbb{R},\mathbb{R})$, telles que $\mathbb{E}(\varepsilon) = 0$ et, pour tout $t < 0$, nous avons $\varepsilon(t) = X(t) = 0$.

$(HA2_{FCVM})$ La fonction inconnue $\theta \in L^2(\mathbb{R},\mathbb{R})$ et, pour tout $t < 0$, nous avons $\theta(t) = 0$.

$(HA3_{FCVM})$ Les espérances $\mathbb{E}(\|\varepsilon\|^2_{L_1})$, $\mathbb{E}(\|X\|^2_{L_1})$, $\mathbb{E}(\|\varepsilon\|^2_{L^2})$ et $\mathbb{E}(\|X\|^2_{L^2})$ sont finies.

Ces hypothèses sont peu contraignantes. Remarquons que, dans les hypothèses $(HA1_{FCVM})$ et $(HA2_{FCVM})$, nous supposons que les fonctions θ et X sont nulles jusqu'à $t = 0$. Cela sous-entend que l'expérience commence à $t = 0$. Sous cette hypothèse, il est possible d'écrire le modèle de convolution (4.1) sous la forme

$$Y(t) = \int_{-\infty}^{+\infty} \theta(s)\, X(t-s)ds + \varepsilon(t).$$

Cela permet dans la suite d'utiliser la transformée de Fourier continue pour tranformer une convolution (sur $\mathbb{R}$) en une multiplication.

4.2.3 Du modèle de convolution au modèle de concurrence

Si nous appliquons la tranformée de Fourier continue $\mathcal{F}$ au modèle de convolution (4.1), nous obtenons l'équivalent de ce modèle dans le domaine des fréquences, connu sous le nom de modèle fonctionnel de concurrence, défini pour tout $\xi \in \mathbb{R}$ par

$$\mathcal{Y}(\xi) = \beta(\xi)\mathcal{X}(\xi) + \mathcal{E}(\xi), \tag{4.4}$$

où β est la fonction à estimer du modèle, $\mathcal{X} := \mathcal{F}(X)$, $\mathcal{Y} := \mathcal{F}(Y)$ et $\mathcal{E} := \mathcal{F}(\varepsilon)$ sont les transformées de Fourier respectives de X, Y et ε.

Les hypothèses générales, sur le modèle de concurrence sont les suivantes.

$(HA1_{FCM})$ Les variables $\mathcal{X}$ et $\mathcal{E}$ sont indépendantes, à valeurs dans $C_0 \cap L^2$, avec $\mathbb{E}(\mathcal{E}) = 0$.

$(HA2_{FCM})$ Nous avons $\beta \in C_0 \cap L^2$.

$(HA3_{FCM})$ Les espérances $\mathbb{E}(\|\mathcal{E}\|^2_{C_0})$, $\mathbb{E}(\|\mathcal{X}\|^2_{C_0})$, $\mathbb{E}(\|\mathcal{E}\|^2_{L^2})$ et $\mathbb{E}(\|\mathcal{X}\|^2_{L^2})$ sont finies.

Ces hypothèses sur le modèle de concurrence correspondent aux hypothèses $(HA1_{FCVM})$, $(HA2_{FCVM})$ et $(HA3_{FCVM})$ sur le modèle de convolution. Elles sont vérifiées à partir du moment où les hypothèses $(HA1_{FCVM})$, $(HA2_{FCVM})$ et $(HA3_{FCVM})$ le sont. La preuve repose essentiellement sur les propriétés de la transformée de Fourier (voir Pinsky (2002)).

4.3 Estimation

L'estimation de la fonction θ se fait en trois temps. Le premier est le passage du modèle de convolution au modèle de concurrence présenté dans le paragraphe précédent. Par la suite, nous estimons la fonction β du modèle de concurrence à l'aide d'un estimateur de type ridge. Enfin, nous revenons au modèle de convolution et à l'estimation de θ en appliquant la transformée de Fourier inverse à l'estimateur de la fonction β.

La définition de l'estimateur de β dans le modèle (4.4) est inspirée de l'estimateur introduit par Hoerl (1962) pour la méthode de régularisation ridge pour traiter des problèmes mal posés en régression linéaire multivariée. Soit $\lambda_n > 0$ un paramètre de régularisation, nous définissons l'estimateur ridge fonctionnel de β par

$$\widehat{\beta}_n := \frac{\frac{1}{n}\sum_{i=1}^{n} \mathcal{Y}_i \, \mathcal{X}_i^*}{\frac{1}{n}\sum_{i=1}^{n} |\mathcal{X}_i|^2 + \frac{\lambda_n}{n}}, \tag{4.5}$$

où l'exposant $*$ désigne le conjugué complexe. En régression linéaire multivariée, Hoerl et Kennard (1970) ont prouvé qu'il existe toujours un paramètre de régularisation pour lequel l'estimateur ridge est meilleur que l'estimateur par moindres carrés ordinaires. Une étude asymptotique de l'estimateur ridge dans ce cadre a été faite notamment par Huh et Olkin (1995). Dans le contexte de la régression linéaire fonctionnelle avec sortie scalaire, Hall et Horowitz (2007) utilisent aussi une methode de régularisation de type ridge pour inverser l'opérateur de covariance de $\mathcal{X}$. Leur approche présente deux différences principales avec la notre : la variable réponse que nous considérons est fonctionnelle, et nous n'inversons que les termes diagonaux de l'opérateur de covariance de $\mathcal{X}$.

Finalement, à partir de l'estimateur de β, l'estimateur de θ est défini par

$$\widehat{\theta}_n := \mathcal{F}^{-1}(\widehat{\beta}_n) = \mathcal{F}^{-1}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} \mathcal{Y}_i \, \mathcal{X}_i^*}{\frac{1}{n}\sum_{i=1}^{n} |\mathcal{X}_i|^2 + \frac{\lambda_n}{n}}\right). \tag{4.6}$$

4.4 Résultats de convergence

4.4.1 Résultats sur le modèle de concurrence

Les résultats de consistence de l'estimateur $\widehat{\beta}_n$ de β sont basés sur la décomposition biais-variance que l'on déduit de la définition (4.5) de l'estimateur

$$\widehat{\beta}_n = \beta - \frac{\lambda_n}{n}\left[\frac{\beta}{\frac{1}{n}\sum_{i=1}^{n} |\mathcal{X}_i|^2 + \frac{\lambda_n}{n}}\right] + \frac{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathcal{X}_i^*}{\frac{1}{n}\sum_{i=1}^{n} |\mathcal{X}_i|^2 + \frac{\lambda_n}{n}}. \tag{4.7}$$

De manière classique, la pénalisation introduit un biais mais permet de contrôler la variance (dernier terme de (4.7)) en réalisant un compromis. Comme $\mathbb{E}(\mathcal{X}) = 0$, la partie du dénominateur $\frac{1}{n}\sum_{i=1}^{n} |\mathcal{X}_i(t)|^2$ peut être proche de zéro pour certaines valeurs de t. La pénalisation ($\lambda_n > 0$) prend tout son sens pour empêcher un dénominateur trop petit et par conséquent éviter d'avoir un estimateur trop irrégulier.

Theorème 4.1. *Considérons le modèle fonctionnel de concurrence avec ses hypothèses générales ($HA1_{FCM}$), ($HA2_{FCM}$) et ($HA3_{FCM}$). Soient $(\mathcal{X}_i, \mathcal{Y}_i)_{i=1,\dots,n}$ des réalisations i.i.d. suivant ce modèle. Nous supposons de plus que*

(A1) $\overline{supp(|\beta|)} \subseteq \overline{supp(\mathbb{E}(|\mathcal{X}|))}$,

(A2) $(\lambda_n)_{n\geq 1}$ *vérifie* $\frac{\lambda_n}{n} \to 0$ *et* $\frac{\sqrt{n}}{\lambda_n} \to 0$ *lorsque* $n \to +\infty$.

Sous ces hypothèses, nous avons

$$\lim_{n\to+\infty} \left|\widehat{\beta}_n - \beta\right|_{L^2} = 0 \quad \textit{en probabilité.}$$

L'hypothèse $(A2)$ est classique dans le contexte de la régression ridge. L'hypothèse $(A1)$ spécifie qu'il n'est pas possible d'estimer β en dehors du support de $|\mathcal{X}|$, ce qui est naturel au vu de l'écriture du modèle fonctionnel de concurrence.

Dans la suite, nous donnons un résultat de convergence avec vitesse pour l'estimateur $\widehat{\beta}_n$. Pour cela, nous devons contrôler la forme des fonctions β et $\mathbb{E}(|\mathcal{X}|)$ aux points de la frontière du support de $\mathbb{E}(|\mathcal{X}|)$. En considérant l'ensemble $C_{\beta,\partial\mathcal{X}} := supp(|\beta|) \cap \partial(supp(\mathbb{E}(|\mathcal{X}|)))$, le résultat suivant permet d'obtenir une vitesse de convergence.

Theorème 4.2. *Considérons le modèle fonctionnel de concurrence avec les hypothèses générales* $(HA1_{FCM})$, $(HA2_{FCM})$ *and* $(HA3_{FCM})$. *Nous supposons que l'hypothèse* $(A1)$ *est réalisée, ainsi que les hypothèses suivantes.*

(A3) $\mathbb{E}\left(\left||\mathcal{X}|^2\right|_{L^2}^2\right) < \infty$.

(A4) $\left\|\frac{|\beta|}{\mathbb{E}(|\mathcal{X}|^2)}\mathbb{1}_{\overline{supp(\beta)}\setminus\partial(supp(\mathbb{E}(|X|)))}\right\|_{L^2} < +\infty$.

(A5) Il existe des nombres réels positifs $\alpha > 0$, $M_0, M_1, M_2, L_I > 0$ *tels que pour tout* $p \in C_{\beta,\partial\mathcal{X}}$, *il existe un voisinage ouvert* $J_p \subset supp(|\beta|)$ *de longueur* $m(J_p) < L_I$ *pour lequel les assertions suivantes sont réalisées*

(a) Pour tout $t \in J_p$, $\mathbb{E}(|\mathcal{X}|^2(t)) \geq |t-p|^\alpha$, *et*

$$\left\|\frac{1}{\mathbb{E}(|\mathcal{X}|^2)}\right\|_{L^2(J_p\setminus\{p\})} \leq M_0,$$

(b) $\sum_{p\in C_{\beta,\partial\mathcal{X}}} \|\beta\|^2_{C_0(J_p)} < M_1$,

(c) $\frac{|\beta|}{\mathbb{E}(|\mathcal{X}|^2)}\mathbb{1}_{supp(|\beta|)\setminus J} < M_2$, *où* $J := \bigcup_{p\in C_{\beta,\partial\mathcal{X}}} J_p$.

(A6) Pour $n \geq 1$,

$$\lambda_n := n^{1-\frac{1}{4\alpha+2}},$$

où $\alpha > 0$ *provient de l'hypothèse* $(A5)$.

Alors

$$\left\|\widehat{\beta}_n - \beta\right\|_{L^2} = O_P\left(n^{-\gamma}\right),$$

où $\gamma := \min\left[\frac{1}{2(2\alpha+1)}, \frac{1}{2} - \frac{1}{2(2\alpha+1)}\right]$ *et* $n^{-\gamma} = \max\left[\frac{\lambda_n}{n}, \frac{\sqrt{n}}{\lambda_n}\right]$.

Corollaire 4.3. *Sous les hypothèses du Théorème 4.2, si* $\alpha < 1/2$*, nous avons*

$$\left\|\widehat{\beta}_n - \beta\right\|_{L^2} = O_P\left(n^{-\gamma}\right),$$

où $\gamma := \alpha/(2\alpha+1) < 1/4$.

L'hypothèse $(A3)$ est classique et permet d'appliquer le TCL sur le dénominateur de $\widehat{\beta}_n$. L'hypothèse $(A4)$ est imposée pour pouvoir borner dans L^2 le terme $\left[\frac{\beta}{\frac{1}{n}\sum_{i=1}^n |\mathcal{X}_i|^2 + \frac{\lambda_n}{n}}\right]$ de la décomposition (4.7).

L'hypothèse $(A5a)$ requiert que, au voisinage des points $p \in C_{\beta,\partial\mathcal{X}}$, la fonction $\mathbb{E}(|\mathcal{X}|^2)$ tend vers zéro plus lentement qu'un polynôme de degré α, ce qui implique que le terme $\left[\frac{\beta}{\frac{1}{n}\sum_{i=1}^n |\mathcal{X}_i|^2 + \frac{\lambda_n}{n}}\right]$ dans la décomposition (4.7) se comporte comme $\frac{\beta}{\mathbb{E}(|\mathcal{X}|^2)}$ et détermine la vitesse de convergence.

La quantité $\mathbb{E}(|\mathcal{X}|^2)$ est cruciale dans le comportement asymptotique de l'estimateur. En effet, plus cette quantité est proche de zéro, plus le problème est mal posé. L'hypothèse $(A5a)$ mesure cela à travers le degré polynômial α.

Les hypothèses $(A5b)$ et $(A5c)$ permettent de contrôler les queues de β et $|\mathcal{X}|$ à l'infini. Ces hypothèses ne sont nécessaires que lorsque $card(C_{\beta,\partial\mathcal{X}}) = +\infty$. Remarquons que l'ensemble $C_{\beta,\partial\mathcal{X}}$ est toujours dénombrable.

Enfin, l'hypothèse $(A6)$ remplace l'hypothèse $(A2)$ du Théorème 4.1, en étant plus forte. La vitesse de convergence dépend directement du comportement de $\frac{\beta}{\mathbb{E}(|\mathcal{X}|^2)}$ autour des points de $C_{\beta,\partial\mathcal{X}}$, qui dépend de α.

D'autres résultats de convergence peuvent être obtenus, en particulier sur l'erreur quadratique moyenne de l'estimateur et sur la construction d'intervalles de confiance de β. Ces résultats sont donnés dans Manrique et al. (2018).

4.4.2 Résultats sur le modèle de convolution

Une fois établis les résultats de convergence de l'estimateur $\widehat{\beta}_n$ vers β, la propriété d'isométrie de la transformée de Fourier continue $\mathcal{F}$ permet d'obtenir les résultats de convergence de $\widehat{\theta}_n$ vers θ. Nous commençons par donner un résultat de consistence avant d'énoncer un résultat qui précise une vitesse de convergence.

Theorème 4.4. *Considérons le modèle fonctionnel de convolution avec les hypothèses générales* $(HA1_{FCVM})$*,* $(HA2_{FCVM})$ *et* $(HA3_{FCVM})$*. Soit* $(X_i, Y_i)_{i=1,\dots,n}$ *un échantillon i.i.d. de même loi que* (X, Y) *issu de ce modèle. Nous supposons de plus que*

(A1) $\overline{supp(|\mathcal{F}(\theta)|)} \subseteq \overline{supp\left(\mathbb{E}\left(|\mathcal{F}(X)|\right)\right)}$,

(A2) $(\lambda_n)_{n\geq 1}$ *vérifie* $\frac{\lambda_n}{n} \to 0$ *et* $\frac{\sqrt{n}}{\lambda_n} \to 0$ *lorsque* $n \to +\infty$.

Alors

$$\lim_{n\to+\infty} \left\|\widehat{\theta}_n - \theta\right\|_{L^2} = 0 \quad \textit{en probabilité.}$$

Nous retrouvons l'hypothèse classique $(A2)$, comme dans le Théorème 4.1. L'hypothèse $(A1)$ spécifie qu'il n'est pas possible d'estimer $\mathcal{F}(\theta)$ en dehors du support de $\mathbb{E}(|\mathcal{F}(X)|)$. Dans les intervalles où $\mathbb{E}(|\mathcal{X}|) = 0$, le modèle (4.4) devient $\mathcal{Y} = \varepsilon$ et aucune estimation de β n'est possible.

En considérant l'ensemble $C_{\theta,\partial X} := supp(|\theta|) \cap \partial(supp(\mathbb{E}(|X|)))$, nous donnons maintenant un résultat de vitesse de convergence.

Theorème 4.5. *Considérons le modèle fonctionnel de convolution avec les hypothèses générales* $(HA1_{FCVM})$, $(HA2_{FCVM})$ *and* $(HA3_{FCVM})$. *Nous supposons que l'hypothèse* $(A1)$ *est réalisée, ainsi que les hypothèses suivantes.*

(A3) $\mathbb{E}\left(\left||\mathcal{X}|^2\right|_{L^2}^2\right) < \infty$.

(A4) $\left\|\frac{|\mathcal{F}(\theta)|}{\mathbb{E}(|\mathcal{F}(X)|^2)} \mathbb{1}_{\overline{supp(\mathcal{F}(\theta))}\setminus\partial(supp(\mathbb{E}(|\mathcal{F}(X)|)))}\right\|_{L^2} < +\infty$.

(A5) Il existe des nombres réels positifs $\alpha > 0$, $M_0, M_1, M_2 > 0$ *tels que pour tout* $p \in C_{\theta,\partial X}$, *il existe un voisinage ouvert* $J_p \subset supp(|\mathcal{F}(\theta)|)$ *pour lequel les assertions suivantes sont réalisées*

(a) Pour tout $t \in J_p$, $\mathbb{E}(|\mathcal{F}(X)|^2(t)) \geq |t-p|^\alpha$, *et*

$$\left\|\frac{1}{\mathbb{E}(|\mathcal{F}(X)|^2)}\right\|_{L^2(J_p\setminus\{p\})} \leq M_0,$$

(b) $\sum_{p\in C_{\theta,\partial X}} \|\theta\|^2_{C_0(J_p)} < M_1$,

(c) $\frac{|\mathcal{F}(\theta)|}{\mathbb{E}(|\mathcal{F}(X)|^2)} \mathbb{1}_{supp(|\mathcal{F}(\theta)|)\setminus J} < M_2$, *où* $J := \bigcup_{p\in C_{\theta,\partial X}} J_p$.

(A6) Pour $n \geq 1$,

$$\lambda_n := n^{1-\frac{1}{4\alpha+2}},$$

où $\alpha > 0$ *provient de l'hypothèse* $(A5)$.

Alors

$$\left\|\widehat{\theta}_n - \theta\right\|_{L^2} = O_P\left(n^{-\gamma}\right),$$

où $\gamma := \min\left[\frac{1}{2(2\alpha+1)}, \frac{1}{2} - \frac{1}{2(2\alpha+1)}\right]$ *et* $n^{-\gamma} = \max\left[\frac{\lambda_n}{n}, \frac{\sqrt{n}}{\lambda_n}\right]$.

Comme cela a déjà été dit pour le modèle de concurrence, l'hypothèse $(A3)$ permet d'utiliser le TCL sur le dénominateur de $\widehat{\beta}_n$. L'hypothèse $(A4)$ est une façon de mesurer la régularité de $\mathcal{F}(\theta)$ vis-à-vis de celle de $\mathbb{E}\left[|\mathcal{F}(X)|^2\right]$. L'hypothèse que ce quotient appartient à $L^2 \cap L^\infty$, au moins sur un voisinage de J (défini dans l'hypothèse $(A5)$), peut être vue comme une condition de plus grande régularité de θ par rapport à $\mathbb{E}(X)$. Par exemple, cela implique que $\mathcal{F}(\theta)$ décroit vers zéro à l'infini plus rapidement que $\mathbb{E}\left[|\mathcal{F}(X)|^2\right]$, ce qui se traduit par le fait que θ est plus lisse que $\mathbb{E}(X)$, c'est-à-dire qu'il existe des entiers $r_1, r_2 \in \mathbb{N}$ tels que θ est de classe r_1 et $\mathbb{E}(X)$ est de classe r_2 avec $r_1 > r_2$ (voir Pinsky (2002)). Ce type de conditions de régularité est utilisé fréquemment dans les problèmes de régression fonctionnelle, où le coefficient fonctionnel du modèle est supposé être plus régulier que la variable explicative (voir par exemple Cardot et al. (2003)).

Il est possible d'obtenir d'autres résultats de convergence, sous des hypothèses un peu plus fortes sur le quotient $\mathcal{F}(\theta)/\mathbb{E}\left[|\mathcal{F}(X)|^2\right]$ (voir Manrique et al. (2019)). La vitesse de convergence devient alors indépendante du comportement de $\mathbb{E}\left[|\mathcal{F}(X)|^2\right]$ autour des points frontière de $C_{\theta,\partial X}$.

4.5 Simulations

Nous présentons dans cette section des expérimentations numériques pour illustrer le comportement de notre estimateur par rapport à d'autres issus de la littérature. Nous commençons par aborder le problème du choix du paramètre de régularisation.

4.5.1 Choix du paramètre de régularisation

Nous présentons une procédure de choix du paramètre de régularisation λ_n pour un échantillon donné $(X_i, Y_i)_{i=1,\dots,n}$, basée sur le critère de validation croisée Leave-One-Out (LOOCV). Sa définition, tirée par exemple de Febrero-Bande et de la Fuente (2012) ou Hall et Hosseini-Nasab (2006), est la suivante

$$LOOCV(\lambda_n) = \frac{1}{n}\sum_{i=1}^{n}\left\|Y_i - \int_0^{\cdot} \widehat{\theta}_n^{(-i)}(s)\, X_i(.-s)\, ds\right\|^2,$$

où $\widehat{\theta}_n^{(-i)}$ est calculé à partir de l'échantillon $(X_j, Y_j)_{j=1,\dots,i-1,i+1,\dots,n}$. La méthode de sélection consiste à choisir la valeur λ_n qui minimise le critère $LOOCV$. Ce critère étant connu pour être coûteux en temps de calcul, nous donnons dans la Proposition 4.6 une façon de calculer le critère LOOCV en calculant une seule régression au lieu de n (voir Green et Silverman (1993)) directement dans le domaine des fréquences.

Proposition 4.6. *Nous avons*

$$LOOCV(\lambda_n) = \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{\mathcal{Y}_i - \widehat{\beta}_n\,\mathcal{X}_i}{1-A_{i,i}}\right\|^2, \tag{4.8}$$

où $A_{i,i} \in L^2$ *est défini par* $A_{i,i} = |\mathcal{X}_i|^2/(\sum_{j=1}^n |\mathcal{X}_j|^2 + \lambda_n)$.

4.5.2 Implémentation numérique

Dans ce paragraphe, nous mentionnons les points principaux permettant de réaliser l'implémentation numérique de l'estimateur. Tout d'abord, notons que les fonctions sont évaluées en un certain nombre de points de discrétisation $t_1, \dots, t_p$, permettant ainsi d'approcher les intégrales considérées à l'aide de la méthode des rectangles. En particulier, la convolution entre deux fonctions sera approchée par la convolution discrète entre les deux vecteurs correspondant aux valeurs des deux fonctions en chaque point de discrétisation.

De plus, nous devons utiliser une approximation de la transformée de Fourier continue et de son inverse (voir (4.5) et (4.6)). Cela se fait à l'aide de la transformée de Fourier discrète (voir Kammler (2008), Bloomfield (2004)), définie de la façon suivante. Pour toute fonction f, on note $\widetilde{\boldsymbol{f}}$ le vecteur dont les coordonnées sont $f(t1), \dots, f(t_p)$. On note alors $\mathcal{F}_d$ la transformée de Fourier discrète définie par

$$\begin{array}{rccc} \mathcal{F}_d\colon & \mathbb{C}^p & \to & \mathbb{C}^p \\ & \widetilde{\boldsymbol{f}} = (f(t_1), \dots, f(t_p)) & \mapsto & (\mathcal{F}_d(f)(1), \dots, \mathcal{F}_d(f)(p)), \end{array}$$

où, pour tout $\ell = 1, \dots, p$

$$\mathcal{F}_d(f)(\ell) = \frac{1}{p}\sum_{r=1}^{p} f(t_r)\omega^{(r-1)\ell} \in \mathbb{C}, \tag{4.9}$$

avec $\omega = e^{-2\pi i/p}$. En utilisant des notations matricielles, nous pouvons écrire

$$\mathcal{F}_d(f) = \frac{1}{p}\boldsymbol{\Omega}_p\widetilde{\boldsymbol{f}} \in \mathbb{C}^p, \tag{4.10}$$

où $\boldsymbol{\Omega}_p$ est la matrice de terme général $(\boldsymbol{\Omega}_p)_{ij} = \omega^{(i-1)(j-1)}$. De plus, nous avons

$$\mathcal{F}_d^{-1} = \boldsymbol{\Omega}_p^*, \tag{4.11}$$

ce qui correspond à la transposée conjuguée de la matrice $\mathbf{\Omega}_p$.

Nous pouvons alors définir l'estimateur de θ en pratique sur la grille $t_1, \ldots, t_p$ par

$$\widetilde{\boldsymbol{\theta}}_n = \frac{1}{p}\mathbf{\Omega}_p^{-1}\left[\frac{\sum_{i=1}^n \mathbf{\Omega}_p\widetilde{\boldsymbol{Y_i}}\left(\mathbf{\Omega}_p\widetilde{\boldsymbol{X_i}}\right)^*}{\sum_{i=1}^n\left|\mathbf{\Omega}_p\widetilde{\boldsymbol{X_i}}\right|^2+\lambda_n}\right]. \tag{4.12}$$

L'avantage principal de cet estimateur est qu'il peut être calculé très rapidement, à l'aide de l'algorithme de la transformée de Fourier rapide (FFT, Fast Fourier Transform). Pour plus détails sur cet algorithme, nous renvoyons le lecteur aux travaux Cooley et Tukey (1965) et Hassanieh et al. (2012) par exemple.

4.5.3 Présentation d'autres méthodes d'estimation

À ce jour et à notre connaissance, il existe peu de méthodes d'estimation adaptées au contexte du modèle (4.1). Dans ce qui suit, nous présentons certaines méthodes adaptées à l'estimation de θ dans ce modèle.

4.5.3.1 Déconvolution de Wiener paramétrique (ParWD)

Cette méthode fait partie de la faille des méthodes de traitement du signal (voir Gonzalez et al. (2009)). Pour chaque observation (X_i, Y_i), X_i est vu comme un signal impulsion et Y_i est vu comme le signal observé, pour $i = 1, \ldots, n$. Nous utilisons l'estimateur **ParWD** pour estimer le signal inconnu θ

$$\widehat{\theta}_{wie,i} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(Y_i)\mathcal{F}(X_i)^*}{|\mathcal{F}(X_i)|^2+\alpha}\right). \tag{4.13}$$

De cette manière, nous obtenons n estimateurs de θ. Leur moyenne sera l'estimateur final de θ dans (4.1), c'est-à-dire $\widehat{\theta}_{ParWD} = \frac{1}{n}\sum_{i=1}^n \widehat{\theta}_{wie,i}$.

Afin de calibrer le paramètre α, nous utilisons le critère LOOCV. Notons que ce paramètre α remplace le ratio signal-sur-bruit $\mathbb{E}\left[|\mathcal{F}(\varepsilon)|^2\right] / |\mathcal{F}(\theta)|^2$ de la méthode originale de déconvolution de Wiener (voir Gonzalez et al. (2009)).

En comparant cet estimateur avec le nôtre, nous pouvons voir que les deux sont reliés. La différence est que la moyenne est faite sur chaque estimateur calculé sur une observation pour l'estimateur ParWD, alors que nous commençons par calculer la moyenne sur les observations dans le domaine des fréquences pour estimer $\mathcal{F}(\theta)$ et ensuite la transformée de Fourier inverse est utilisée pour revenir dans le domaine temporel. Nous verrons dans la suite que les deux estimateurs ont un comportement similaire.

4.5.3.2 Décomposition en valeurs singulières (SVD) ou régularisation de Tikhonov (Tik)

La convolution peut être vue comme une multiplication matricielle, et par conséquent, la déconvolution peut être vue comme un problème d'inversion matricielle. Nous allons ainsi considérer deux techniques bien connues d'inversion matricielle, la décomposition en valeurs singulières (SVD, Singular Value Decomposition) et la régularisation de Tikhonov.

Étant donné que la solution est fortement sensible au bruit, nous devons éliminer le bruit le plus possible avant inversion. Par conséquent, nous commençons par calculer la moyenne sur toutes les observations, de façon à obtenir, pout tout t

$$\overline{Y}(t) = \int_0^t \theta(s)\overline{X}(t-s)ds + \overline{\varepsilon}(t), \tag{4.14}$$

où $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ et $\overline{\varepsilon} = \frac{1}{n}\sum_{i=1}^n \varepsilon_i$. Notons que cette méthode diffère de la méthode précédente ParWD, où la moyenne était calculée après estimation.

Ensuite, nous approximons cette équation intégrale sous forme matricielle en utilisant la méthode des rectangles sur une grille $t_1, \ldots, t_p$ de points de discrétisation. Nous obtenons

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{M_X}\widetilde{\boldsymbol{\theta}} + \widetilde{\boldsymbol{\varepsilon}},$$

où $\widetilde{\boldsymbol{Y}} = (Y(t_1), \ldots, Y(t_p))^T$, $\widetilde{\boldsymbol{\theta}} = (\theta(t_1), \ldots, \theta(t_p))^T$, $\widetilde{\boldsymbol{\varepsilon}} = (\varepsilon(t_1), \ldots, \varepsilon(t_p))^T$ et $\boldsymbol{M_X}$ est la matrice triangulaire inférieure qui approxime la convolution en ces points de discrétisation, en d'autres termes

$$\boldsymbol{M_X} = \begin{pmatrix} \overline{X}(t_1) & 0 & 0 & \ldots & 0 \\ \overline{X}(t_2) & \overline{X}(t_1) & 0 & \ldots & 0 \\ \overline{X}(t_3) & \overline{X}(t_1) & \overline{X}(t_1) & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \overline{X}(t_p) & \overline{X}(t_{p-1}) & \overline{X}(t_{p-2}) & \ldots & \overline{X}(t_1) \end{pmatrix}.$$

Nous considérons la SVD de $\boldsymbol{M_X}$, soit $\boldsymbol{M_X} = \boldsymbol{USV}^T$ où $\boldsymbol{S}$ est la matrice diagonale avec les valeurs singulières de $\boldsymbol{M_X}$ sur la diagonale (soit les racines carrées des valeurs propres de ${\boldsymbol{M_X}}^T\boldsymbol{M_X}$) et $\boldsymbol{U}$ et $\boldsymbol{V}$ sont des matrices orthogonales.

L'estimateur de θ par la méthode SVD est défini par

$$\widehat{\theta}_{SVD} = \boldsymbol{V}\boldsymbol{S}_k^+\boldsymbol{U}^T\widetilde{\boldsymbol{Y}},$$

où $\boldsymbol{S}_k$ est la matrice diagonale avec les mêmes k non nuls premiers éléments que $\boldsymbol{S}$ (et zéro ailleurs), et $\boldsymbol{S}_k^+$ est le pseudo-inverse de $\boldsymbol{S}_k$, obtenu en remplaçant les éléments non

nuls de la diagonale de $\boldsymbol{S}$ par leurs inverses. Ici, le paramètre de régularisation est la dimension k.

L'estimateur de θ par la méthode de Tikhonov est défini par

$$\widehat{\theta}_{Tik} = \boldsymbol{V}\boldsymbol{S}(\boldsymbol{S}^2 + \rho\boldsymbol{I})^{-1}\boldsymbol{U}^T\widetilde{\boldsymbol{Y}},$$

où ρ est le paramètre de régularisation.

Pour calibrer les paramètres de ces deux estimateurs, nous utilisons la validation croisée 10-fold (voir Seni et Elder (2010)).

4.5.3.3 Estimateur de Laplace (Lap)

Nous utilisons ici une version adaptée de l'estimateur de Laplace introduit dans Comte et al. (2017), noté $\widehat{\theta}_{Lap}$. Nous commençons par calculer la moyenne de toutes les réalisations pour éliminer le bruit, étant donné que cet estimateur est construit pour résoudre le problème de déconvolution avec une seule observation. Nous obtenons ainsi, pour $j = 1, \ldots, p$

$$\overline{Y}(t_j) = \int_0^{t_j} \theta(s)\overline{X}(t_j - s)ds + \overline{\varepsilon}(t_j) \tag{4.15}$$

où $t_1, \ldots, t_p$ sont les p points de mesure.

Dans Comte et al. (2017), les auteurs estiment θ à l'aide de fonctions de Laguerre, définies pour $k \in \mathbb{N}$, $t \geq 0$ et $a > 0$ fixé, par

$$\phi_k(t) = \sqrt{2a}e^{-at}\left(\sum_{j=0}^{k}(-1)^j\binom{k}{j}\frac{t^j}{j!}\right).$$

Ces fonctions sont utilisées en tant que base orthonomée de $L^2(\mathbb{R}_+, \mathbb{R})$ pour transformer l'équation (4.15) en un système infini d'équations linéaires dont les coefficients sont exprimés dans la base de Laguerre. La convolution de deux fonctions de Laguerre s'exprime de façon simple, pour $k, \ell \geq 0$ par

$$\int_0^t \phi_k(s)\phi_\ell(t - s)ds = (2a)^{-1/2}[\phi_{k+\ell}(t) - \phi_{k+\ell+1}(t)].$$

Ainsi, le système se simplifie en un système infini triangulaire inférieur d'équations linéaires. Le sous-système fini des M premières équations est résolu, permettant de calculer les M premiers coefficients de $\widehat{\theta}_{Lap}$ dans la base de Laguerre. L'implémentation numérique de cet estimateur peut être trouvée dans le package R nommé LaplaceDeconv.

Échantillon	FFDE	ParWD	SVD	Tik	Lap
n=70	**0.152**	10.097	9.587	3.735	3.071
n=400	**0.705**	225.313	14.469	6.107	4.593

Table 4.1 Temps de calcul (en secondes) des estimateurs.

4.5.4 Exemple de simulations

Dans cette partie, nous présentons sur un exemple simulé le comportement de notre estimateur ainsi que des estimateurs présentés dans le paragraphe précedent. Des simulations plus poussées sont données dans Manrique et al. (2019).

Les données ont été simulées sur l'intervalle $[0;1]$ (soit $T = 1$), discrétisé en $p = 100$ points d'observation $t_j = (j-1)/100$, avec $j = 1, \ldots, 100$. Nous avons simulé chaque courbe X comme un pont Brownien sur l'intervalle $[0;0.5]$ et identiquement nulle sur l'intervalle $[0.5;1]$. Pour la fonction θ, nous avons considéré la fonction définie par $\theta(t) = (1-4t^2)$ sur l'intervalle $[0;0.5]$ et identiquement nulle sur l'intervalle $[0.5;1]$. Le bruit ε est un bruit blanc gaussien d'écart-type σ choisi tel que le rapport signal-sur-bruit soit égal à 10 (soit 10% de bruit). Nous évaluons les performances des estimateurs pour des tailles d'échantillon $n = 70$ et $n = 400$. L'erreur d'estimation sera mesurée à l'aide des deux critères suivants : l'erreur moyenne absolue (MADE, Mean Absolute Deviation Error) et l'erreur moyenne quadratique (WASE, Weighted Average Squared Error), tel qu'ils ont été définis dans Sentürk et Müller (2010),

$$MADE = \frac{1}{T}\left[\frac{\int_0^T \left|\theta(t) - \widehat{\theta}(t)\right| dt}{\text{range}(\theta)}\right], \qquad WASE = \frac{1}{T}\left[\frac{\int_0^T \left|\theta(t) - \widehat{\theta}(t)\right|^2 dt}{\text{range}^2(\theta)}\right],$$

où $\text{range}(\theta) = \max_{t\in I}\theta(t) - \min_{t\in I}\theta(t)$.

En premier lieu, nous donnons les temps de calcul (en secondes) pour une simulation dans la Table 4.1. Les calculs ont été réalisés avec R, sur une machine 2.9 GHz x 4 Intel Core i7-3520M. Nous voyons que l'estimateur FFDE est plus rapide que les autres.

La Figure 4.1 montre la vraie fonction θ et les cinq estimateurs moyens calculés sur $N = 100$ simulations. Les meilleurs estimateurs sont **FFDE** et **parWD**, tous deux sont proches l'un de l'autre. Nous pouvons noter que **FFDE** a des difficultés d'estimation près des bords de l'intervalle. Par ailleurs, **SVD** et **Tik** ont très variations assez importantes sur l'intervalle d'estimation, tandis que **Lap** a des difficultés d'estimation en ce qui concerne la partie quadratique de θ sur le début de l'intervalle d'estimation. Enfin, les estimateurs s'améliorent lorsque la taille de l'échantillon augmente, notamment **FFDE**.

Plus précisément, nous pouvons voir dans la Table 4.2 et sur les boxplots de la

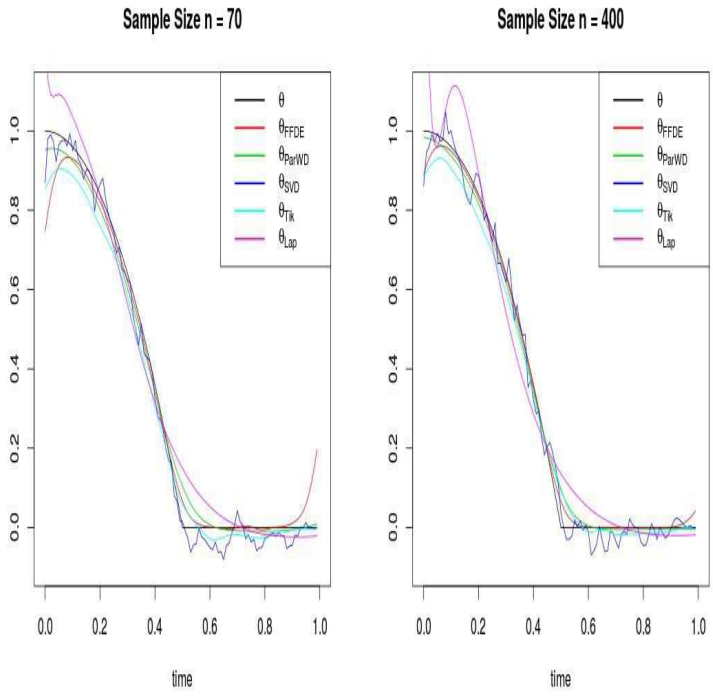

Figure 4.1 Vraie fonction θ comparée aux estimations moyennes des 5 estimateurs pour $N = 100$ simulations.

	MADE	WASE
n = 70	moyenne (écart-type)	moyenne (écart-type)
FFDE	**0.04120 (0.00895)**	**0.00400 (0.00212)**
ParWD	**0.03020 (0.00657)**	**0.00157 (0.00062)**
SVD	0.16240 (0.15467)	0.08356 (0.16906)
Tik	0.08797 (0.04836)	0.01573 (0.01764)
Lap	0.16427 (0.11468)	0.10468 (0.30549)
n = 400	moyenne (écart-type)	moyenne (écart-type)
FFDE	**0.01273 (0.00151)**	**0.00044 (0.00013)**
ParWD	0.02010 (0.00342)	0.00076 (0.00024)
SVD	0.16313 (0.18566)	0.10284 (0.27954)
Tik	0.07641 (0.03702)	0.01120 (0.01170)
Lap	0.18968 (0.13129)	0.15172 (0.28369)

Table 4.2 Moyennes et écart-types des deux critères MADE et WASE, calculés sur $N = 100$ simulations avec des tailles d'échantillon $n = 70$ et $n = 400$.

Figure 4.2 que **FFDE** et **ParWD** sont les meilleurs estimateurs, tandis que **SVD** est le moins bon.

Dans ce cadre de simulations, nous remarquons que **FFDE** et **ParWD** se comportent bien dans le cas où $\mathbb{E}(X) = 0$, parce qu'ils utilisent l'algorithme FFT pour déconvoluer directement X et θ, tandis que **SVD** et **Tik** ont de mauvaises performances parce qu'ils ne peuvent pas inverser facilement la matrice $\boldsymbol{M}_X$. De plus, nous remarquons que l'estimateur **Lap** ne s'améliore pas parce que nous l'appliquons sur l'équation moyenne (4.15), qui est presque la même lorsque $n = 70$ et $n = 400$.

4.6 Perspectives

Il semble envisageable de pouvoir généraliser le modèle de convolution (4.1) dans plusieurs directions.

La première d'entre elles serait de considérer le cas où la variable d'intérêt dépend aussi d'une variable qualitative à J modalités (par exemple, une caractéristique comme le sexe de l'individu, un phénotype, ...). Le modèle considéré serait alors

$$Y_j(t) = \int_0^t \theta_j(s)\, X_j(t-s)ds + \varepsilon_j(t), \tag{4.16}$$

pour $j = 1, \ldots, J$. Dans la thèse de Tito Manrique, la méthodologie présentée dans ce chapitre avait été testée sur un jeu de données réelles de croissance de plants de maïs. Les courbes explicatives sont des courbes journalières de déficit de pression de vapeur (variable liée à l'humidité dans l'air) et les courbes réponses sont des vitesses journalières

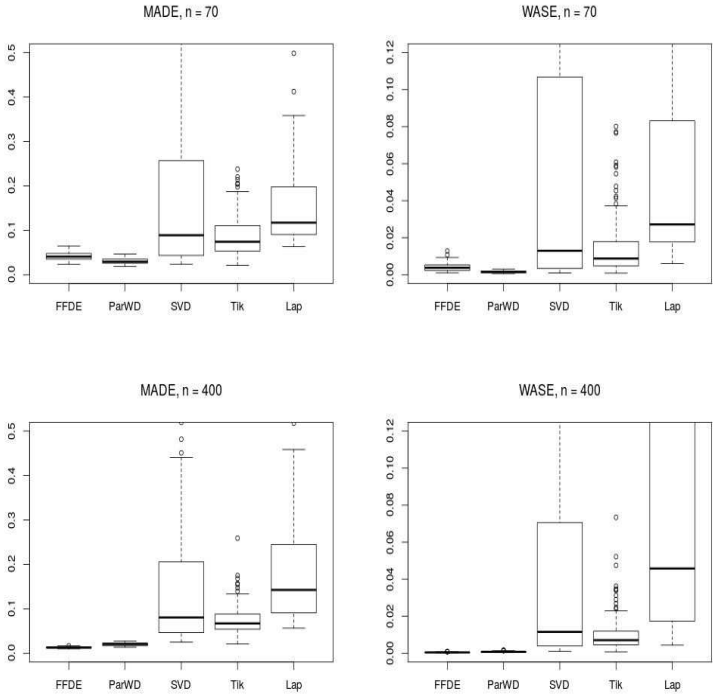

Figure 4.2 Boxplots des deux critères MADE et WASE sur $N = 100$ simulations avec des tailles d'échantillon $n = 70$ et $n = 400$.

de croissance des feuilles. L'idée, non explorée, était de considérer un phénotype comme variable explicative qualitative supplémentaire.

Une autre extension envisagée sur les modèles étudiés dans ce chapitre consiste en un couplage du modèle concurrent et du modèle de convolution. Plus précisément, le modèle considéré est le suivant

$$Y(t) = \int_0^t \theta(s)\, X(t-s)ds + \beta(t)X(t) + \varepsilon(t). \tag{4.17}$$

Intuitivement, ce modèle revient à considérer que la variable réponse est expliquée par le passé de la variable X, ainsi qu'une contribution instantanée. Plusieurs problèmes se posent sur ce type de modèle, le premier étant sur la méthode d'estimation. En effet, la transformée de Fourier présentait l'avantage de transformer la convolution en produit, en revanche elle transformerait aussi le produit βX en convolution. Par ailleurs, considérer une contribution instantanée en tout instant t ne semble pas pertinent, il serait préférable de considérer un certain nombre de points d'impact, comme introduit dans Kneip et al. (2016). Pour adapter ce modèle à une sortie fonctionnelle, il serait peut-être envisageable de considérer des contributions fonctionnelles pour chaque point d'impact

$$Y(t) = \int_0^t \theta(s)\, X(t-s)ds + \sum_{r=1}^{S} \beta_r(t)X(\tau_r) + \varepsilon(t), \tag{4.18}$$

où $\tau_1, \ldots, \tau_S$ sont les points d'impact. L'étude de ce modèle reste un problème ouvert.

Bibliographie

M. Asencio, G. Hooker et H.O. Gao. Functional convolution models. *Statistical Modelling*, 14 :315–335, 2014.

P. Bloomfield. *Fourier analysis of time series : an introduction.* Wiley, 2004.

H. Cardot, F. Ferraty et P. Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, 13 :571–591, 2003.

F. Comte, C.-A. Cuenod, M. Pensky et Y. Rozenholc. Laplace deconvolution on the basis of time domain data and its application to dynamic contrast-enhanced imaging. *Journal of the Royal Statistical Society, Series B*, 79 :69–94, 2017.

J.W. Cooley et K.W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19 :297–301, 1965.

M. Febrero-Bande et M. Oviedo de la Fuente. Statistical computing in functional data analysis : the r package fda.usc. *Journal of Statistical Software*, 51 :1–28, 2012.

R. Gonzalez, R. Woods et S. Eddins. *Digital image processing using Matlab (second edition)*. Gatesmark Publishing, 2009.

P.J. Green et B.W. Silverman. *Nonparametric regression and generalized linear models : a roughness penalty approach*. Chapman and Hall, CRC Press, 1993.

P. Hall et J.L. Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1) :70–91, 2007.

P. Hall et M. Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society : Series B*, 68 :109–126, 2006.

H. Hassanieh, P. Indyk, D. Katabi et E. Price. Nearly optimal sparse fourier transform. *Proceedings of the fourty four annual ACM symposium on theory of computong*, pages 563–578, 2012.

A.E. Hoerl. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58 :54–59, 1962.

A.E. Hoerl et R.W. Kennard. Ridge regression : biased estimation for nonorthogonal problems. *Technometrics*, 12 :55–67, 1970.

M.-H. Huh et I. Olkin. Asymptotic aspects of ordinary ridge regression. *American Journal of Mathematical and Management Sciences*, 15 :239–254, 1995.

D. Kammler. *A first course in Fourier analysis*. Cambridge University Press, 2008.

A. Kneip, D. Poss et P. Sarda. Functional linear regression with points of impact. *Annals of Statistics*, 44 :1–30, 2016.

N. Malfait et J.O. Ramsay. The historical functional linear model. *Canadian Journal of Statistics*, 31 :115–128, 2003.

T. Manrique, C. Crambes et N. Hilgert. Ridge regression for the functional concurrent model. *Electronic Journal of Statistics*, 12 :985–1018, 2018.

T. Manrique, C. Crambes et N. Hilgert. Estimation for the functional convolution model. *Preprint*, 2019.

M. Pinsky. *Introduction to Fourier analysis and wavelets*. American Mathematical Society, 2002.

J.O. Ramsay et B.W. Silverman. *Functional Data Analysis (Second edition)*. Springer-Verlag, New York, 2005.

G. Seni et J. Elder. *Ensemble methods in data mining : improving accuracy through combining predictions.* Synthesis lectures on data mining and knowledge discovery, Morgan and Claypool Publishers, 2010.

D. Sentürk et H.-G. Müller. Functional varying coefficient models for longitudinal data. *Journal of the American Statistical Association*, 105 :1256–1264, 2010.

Chapitre 5

Estimation récursive dans le modèle non-paramétrique fonctionnel

Contents

Les résultats présentés dans ce chapitre sont tirés de Amiri et al. (2014).

5.1 Introduction

Dans ce chapitre, nous nous plaçons dans le cadre du modèle non-paramétrique fonctionnel (voir Ferraty et Vieu (2006)), écrit sous la forme

$$Y = \Psi(X) + \varepsilon, \tag{5.1}$$

où Y et ε sont à valeurs dans $\mathbb{R}$, X est à valeurs dans $L^2(I)$, Ψ est un opérateur de $L^2(I)$ dans $\mathbb{R}$ à estimer sur la base d'un échantillon $(X_i, Y_i)_{i=1,\dots,n}$, et le bruit ε est centré, indépendant de X et tel que $\mathbb{E}\left(\varepsilon^2 | X = x\right) = \sigma_\varepsilon^2(x)$. L'estimation non-paramétrique de Ψ a fait l'objet de nombreux travaux (voir notamment Ferraty et Vieu (2006) et Ferraty et Romain (2010)) tant d'un point de vue théorique que pratique. Plus précisément, si $(X_i, Y_i)_{i=1,\dots,n}$ est un échantillon de n couple de même loi que (X, Y), alors l'estimateur à noyau de Ψ est défini pour tout $\chi \in L^2(I)$ par

$$\Psi_n(\chi) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{\|\chi - X_i\|}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\|\chi - X_i\|}{h}\right)}, \tag{5.2}$$

où K est un noyau et h est une fenêtre. Cet estimateur soulève plusieurs problèmes, comme le choix de la semi-norme $\|.\|$ de l'espace fonctionnel, ou encore le choix de la largeur de fenêtre h. Comme cela avait été fait par Amiri (2012) dans le cadre d'une variable explicative réelle ou multivariée, nous proposons dans ce cadre fonctionnel un estimateur utilisant une suite de fenêtres permettant de calculer cet estimateur de façon récursive (voir aussi les travaux antérieurs Devroye et Wagner (1980) et Ahmad et Lin (1976)). Cet estimateur récursif, en plus d'avoir de bonnes propriétés théoriques en termes de convergence est aussi d'un grand intérêt pratique, présentant un gain de temps pour le calcul d'une nouvelle valeur prédite lorsqu'arrive une nouvelle observation de la variable explicative. Récemment, Slaoui (2019) a revisité l'estimation nonparamétrique récursive à l'aide d'une méthode d'approximation stochastique.

5.2 Estimateur récursif

5.2.1 Notations et définitions

Soit $\ell \in [0;1]$ un paramètre réel. Pour $\chi \in L^2(I)$ donné, on définit une famille d'estimateurs récursifs indexée par ℓ par

$$\Psi_n^{[\ell]}(\chi) = \frac{\sum_{i=1}^{n} \frac{Y_i}{F(h_i)^\ell} K\left(\frac{\|\chi - X_i\|}{h_i}\right)}{\sum_{i=1}^{n} \frac{1}{F(h_i)^\ell} K\left(\frac{\|\chi - X_i\|}{h_i}\right)},$$

où (h_n) est une suite de fenêtres et F est la fonction de répartition de la variable $\|\chi - X\|$. Cette famille d'estimateurs de la régression est l'adaptation au cas où la variable explicative est fonctionnelle de l'estimateur défini dans Amiri (2012). L'écriture récursive de cette famille d'estimateurs est donnée par

$$\Psi_{n+1}^{[\ell]}(x) = \frac{\left[\sum_{i=1}^{n} F(h_i)^{1-\ell}\right] \varphi_n^{[\ell]}(\chi) + \left[\sum_{i=1}^{n+1} F(h_i)^{1-\ell}\right] Y_{n+1} K_{n+1}^{[\ell]}\left(\|\chi - X_{n+1}\|\right)}{\left[\sum_{i=1}^{n} F(h_i)^{1-\ell}\right] f_n^{[\ell]}(\chi) + \left[\sum_{i=1}^{n+1} F(h_i)^{1-\ell}\right] K_{n+1}^{[\ell]}\left(\|\chi - X_{n+1}\|\right)},$$

avec

$$\varphi_n^{[\ell]}(\chi) = \frac{\sum_{i=1}^{n} \frac{Y_i}{F(h_i)^\ell} K\left(\frac{\|\chi - X_i\|}{h_i}\right)}{\sum_{i=1}^{n} F(h_i)^{1-\ell}},$$

$$f_n^{[\ell]}(\chi) = \frac{\sum_{i=1}^{n} \frac{1}{F(h_i)^\ell} K\left(\frac{\|\chi - X_i\|}{h_i}\right)}{\sum_{i=1}^{n} F(h_i)^{1-\ell}},$$

$$K_i^{[\ell]}(.) = \frac{1}{F(h_i)^\ell \sum_{j=1}^{i} F(h_j)^{1-\ell}} K\left(\frac{.}{h_i}\right).$$

Dans cette famille d'estimateurs, on peut noter deux cas particuliers : l'estimateur récursif ($\ell = 0$) et semi-récursif ($\ell = 1$). La propriété de récursivité est un avantage évident pour traiter des données séquentielles arrivant en temps réel, notamment sur de grands échantillons, puisque l'arrivée d'une nouvelle observation nécessite à l'estimateur non-récursif d'être recalculé sur tout l'échantillon.

5.2.2 Hypothèses

Afin de présenter les résultats de convergence de l'estimateur récursif, nous présentons les hypothèses que nous supposons.

(H.1) Les opérateurs Ψ et σ_ε^2 sont continus dans un voisinage de χ et $F(0) = 0$. De plus, la fonction $\varphi(t) := \mathbb{E}\left(r(X) - r(\chi)\right| \|X - \chi\| = t)$ est supposée être dérivable en $t = 0$.

(H.2) Le noyau K a pour support le compact $[0;1]$. De plus, sa dérivée K' est continue sur $[0;1]$ et vérifie $K'(s) \leq 0$ pour tout $s \in [0;1]$. Enfin, $K(1) > 0$.

(H.3) Pour tout $s \in [0;1]$, on a $\tau_h(s) := \frac{F(hs)}{F(h)} \to \tau_0(s) < +\infty$ lorsque $h \to 0$.

(H.4) On suppose les conditions suivantes sur la fenêtre :

$$(i) h_n \to 0 \text{ et } nF(h_n) \to +\infty,$$
$$(ii) A_{n,\ell} := \frac{1}{n}\sum_{i=1}^{n} \frac{h_i}{h_n}\left(\frac{F(h_i)}{F(h_n)}\right)^{1-\ell} \to \alpha_\ell < +\infty \text{ lorsque } n \to \infty,$$
$$(iii) \forall r \leq 2,\ B_{n,r} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{F(h_i)}{F(h_n)}\right)^{r} \to \beta_r < +\infty \text{ lorsque } n \to \infty.$$

Les hypothèses (H.1), (H.2) et (H.4) (i) sont classiques en régression nonparamétrique, comme cela est souligné dans Ferraty et al. (2007). Par ailleurs, l'hypothèse (H.3) est cruciale dans le calcul exact des constantes qui apparaissent dans les développements asymptotiques que nous présentons dans la suite (voir Ferraty et al. (2007)). Enfin, les hypothèses (H.4) (ii) et (iii) sont liées au contexte de l'estimateur récursif et sont similaires à celles utilisées dans le cadre multivarié. Notons que la fonction F joue un rôle crucial dans nos résultats. Sa limite en zéro, pour un χ fixé est connue sous le nom de probabilité de petites boules. Par exemple, si X est un processus fractal, les probabilités de petites boules vérifient $F(h) \sim Ch^{\kappa}$ pour certains C et κ en prenant comme semi-norme une norme dans L^p, une norme de Sobolev ou une norme de Besov, ce qui permet de vérifier les hypothèses précédentes.

5.2.3 Résultats

Comme dans Ferraty et al. (2007), nous considérons

$$M_0 = K(1) - \int_0^1 (sK(s))'\tau_0(s)\, ds,$$
$$M_1 = K(1) - \int_0^1 K'(s)\tau_0(s)\, ds,$$
$$M_2 = K(1)^2 - \int_0^1 (K(s)^2)'\tau_0(s)\, ds.$$

Nous sommes maintenant en mesure d'énoncer un résultat de convergence en moyenne quadratique de l'estimateur nonparamétrique fonctionnel récursif. Les preuves sont données dans Amiri et al. (2014).

Theorème 5.1. *Sous les hypothèses (H.1)-(H.4), nous avons*

$$\mathbb{E}\left(\Psi_n^{[\ell]}(\chi)\right) - \Psi(\chi) = \varphi'(0)\frac{\alpha_\ell}{\beta_{1-\ell}}\frac{M_0}{M_1}h_n(1+o(1)) + O\left(\frac{1}{nF(h_n)}\right),$$
$$\mathbb{V}\left(\Psi_n^{[\ell]}(x)\right) = \frac{\beta_{1-2\ell}}{\beta_{1-\ell}^2}\frac{M_2}{M_1^2}\sigma_\varepsilon^2(\chi)\frac{1}{nF(h_n)}(1+o(1)).$$

Ce résultat étend celui obtenu dans Ferraty et al. (2007) au cas des estimateurs récursifs. En écrivant une décomposition classique de l'erreur quadratique moyenne entre le biais au carré et la variance, nous obtenons de façon immédiate le résultat suivant, pour certains types de processus.

Corollaire 5.2. *Sous les hypothèses du Théorème 5.1, si X est un processus tel que $F(t) \sim c_X t^\kappa$ lorsque $t \to 0$ (avec $\kappa > 0$), alors en prenant $h_n = An^{-1/(\kappa+2)}$, nous obtenons*

$$\mathbb{E}\left[\left(\Psi_n^{[\ell]}(\chi) - \Psi(\chi)\right)^2\right] \sim_{n\to+\infty} \left[\frac{\beta_{1-2\ell}}{\beta_{1-\ell}^2}\frac{M_2\sigma_\varepsilon^2(\chi)}{Ac_X M_1^2} + \frac{\alpha_\ell^2}{\beta_{1-\ell}^2}\frac{A^2M_0^2\varphi'(0)^2}{M_1^2}\right] n^{-\frac{2}{\kappa+2}}.$$

Un résultat similaire est obtenu dans Bosq et Cheze-Payaud (1999) pour l'estimateur de Nadaraya-Watson dans un cadre classique.

Nous allons à présent donner un résultat de convergence presque sûre. Pour cela, nous utilisons les hypothèses supplémentaires suivantes.

(H.5) Il existe $\lambda, \mu > 0$ tels que $\mathbb{E}\left(e^{\lambda|Y|^\mu}\right) < +\infty$.

(H.6) Nous supposons que

$$\begin{aligned}
&(i) \lim_{n\to+\infty} \frac{\ln F(h_n)}{\ln n} < +\infty,\\
&(ii) \lim_{n\to+\infty} \frac{nF(h_n)(\ln n)^{-1-2/\mu}}{(\ln\ln n)^{2(\alpha+1)}} = +\infty \text{ pour tout } \alpha > 0,\\
&(iii) \lim_{n\to+\infty} F(h_n)(\ln n)^{2/\mu} = 0.
\end{aligned}$$

L'hypothèse (H.5) est par exemple clairement vérifiée si Y est borné p.s. et implique que

$$\forall n \geq 2\ \forall p \geq 1\ \mathbb{E}\left(\max_{i=1,\ldots,n} |Y_i|^p\right) = O\left((\ln n)^{p/\mu}\right),$$

qui est une condition que l'on retrouve dans Bosq et Cheze-Payaud (1999). Les hypothèses (H.6) (i) et (ii) sont vérifiées dès que X est un processus fractal, tandis que l'hypothèses (H.6) (iii) n'est nécessaire que si $\mu < 2$. Nous sommes maintenant en mesure d'énoncer le résultat de convergence p.s.

Theorème 5.3. *Sous les hypothèses (H.1)-(H.6), en supposant de plus que* $\lim_{n\to+\infty} nF(h_n)^2 = 0$, *alors*

$$\Psi_n^{[\ell]}(\chi) - \Psi(\chi) \sim_{n\to+\infty} \left[2\beta_{1-2\ell}\sigma_\varepsilon^2(x)M_2\right]^{1/2}\left(\frac{ln\ ln\ n}{nF(h_n)}\right)^{1/2} p.s.$$

Les choix de fenêtres et de petites boules donnés précédemment permettent de satisfaire la condition $\lim_{n\to+\infty} nF(h_n)^2 = 0$. Dans le cas où $\ell = 1$, le résultat du théorème 5.3 étend au cas fonctionnel le résultat de convergence p.s. obtenu par Roussas (1992) sur l'estimateur de Devroye-Wagner. Dans le cas non-paramétrique fonctionnel non-récursif, une vitesse de convergence p.s. de l'ordre de $\left(\frac{\ln n}{nF(h_n)}\right)^{1/2}$ est obtenue dans Ferraty et Vieu (2006). En revanche, le résultat que nous donnons ici fournit l'expression exacte des constantes.

Pour finir, nous allons donner une résultat de normalité asymptotique de l'estimateur récursif. Nous considérons l'hypothèse additionnelle suivante.

(H.7) Pour tout $\delta > 0$, on a $\lim_{n\to+\infty} \frac{(\ln n)^\delta}{(nF(h_n))^{1/2}} = 0$.

Theorème 5.4. *Sous les hypothèses (H.1)-(H.5) et (H.7), s'il existe une constante C telle que $\lim_{n\to+\infty} h_n(nF(h_n))^{1/2} = C$, alors*

$$(nF(h_n))^{1/2}\left(\Psi_n^{[\ell]}(\chi) - \Psi(\chi)\right) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\left(C\frac{\alpha_\ell}{\beta_{1-\ell}}\frac{M_0}{M_1}\varphi'(0), \frac{\beta_{1-2\ell}}{\beta_{1-\ell}^2}\frac{M_2}{M_1^2}\sigma_\varepsilon^2(\chi)\right).$$

Ce résultat étend le résultat de Ferraty et al. (2007) au cas récursif. Notons que, pour les choix de fenêtres et de petites boules précédents, on a $\frac{\beta_{1-2\ell}}{\beta_{1-\ell}^2} < 1$. Ainsi, la variance asymptotique de l'estimateur récursif est plus petite que la variance asymptotique de l'estimateur non récursif. De ce résultat, une bande de confiance asymptotique peut être construite (voir Amiri et al. (2014)).

5.3 Simulations

Dans cette section, nous nous intéressons à la mise en œuvre et aux performances pratiques de l'estimateur récursif, en particulier par rapport à l'estimateur non récursif. Le modèle étudié dans cette simulation est le suivant. Les courbes $X_1, \ldots, X_n$, avec $n = 100$, sont des mouvements Browniens sur l'intervalle $[0; 1]$ mesurés en $p = 100$ points de discrétisation équidistants. L'opérateur Ψ est défini par

$$r(\chi) = \int_0^1 \chi(s)^2\, ds.$$

L'erreur ε est simulée suivant une loi normale centrée et d'écart-type 0.1. Les simulations ont été répétées $N = 500$ fois afin d'étudier l'erreur de prédiction de la réponse pour une nouvelle courbe χ, elle aussi simulée suivant un mouvement Brownien sur $[0; 1]$.

Plusieurs paramètres interviennent dans la construction de l'estimateur : le choix de la semi-norme $\|.\|$, la suite de fenêtres (h_n), le noyau K, et le paramètre ℓ de récursivité.

Le choix du noyau est moins crucial et n'a qu'une influence restreinte dans l'estimation, nous nous limitons ainsi à l'utilisation du noyau quadratique défini par $K(u) = (1 - u^2)\mathbb{1}_{[0;1]}(u)$. En effet, même si ce noyau ne vérifie pas l'hypothèse $K(1) > 0$ demandée pour les résultats asymtotiques, ce noyau est très facile à implémenter et se comporte bien en pratique.

5.3.1 Choix de la fenêtre

Dans cette simulation, nous choisissons comme semi-norme celle basée sur l'analyse en composantes principales des courbes avec 3 composantes principales (voir Besse et al. (1997)), tandis que le paramètre ℓ est fixé égal à zéro (le choix du paramètre ℓ a une influence minime sur le comportement de l'estimateur, voir Amiri et al. (2014)).

Nous choisissons une suite de fenêtres $h_i = C \max_{j=1,\dots,n} \|X_j - \chi\| \, i^{-\nu}$ pour tout $i = 1, \dots, n$, avec $C \in \{0.5, 1, 2, 10\}$ et $\nu \in \left\{\frac{1}{10}, \frac{1}{8}, \frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1\right\}$.

Nous avons également calculé l'estimateur (5.2) introduit par Ferraty et Vieu (2006), en choisissant automatiquement la fenêtre par validation croisée (voir Rachdi et Vieu (2007)). Pour l'estimateur récursif, nous avons considéré le critère de validation croisée (sur les paramètres C et ν)

$$CV(C, \nu) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \Psi_n^{[\ell],[-i]}(X_i) \right)^2,$$

où $\Psi_n^{[\ell],[-i]}$ représente l'estimateur récursif de Ψ construit en utilisant l'échantillon de taille $n - 1$ correspondant à l'échantillon initial privé de l'observation (X_i, Y_i). Nous choisissons alors les valeurs de C et ν qui minimisent le critère $CV(C, \nu)$. La Table 5.1 présente les moyennes et les écart-types de l'erreur quadratique de prédiction ($MSPE$, Mean Square Prediction Error) sur 500 simulations pour les valeurs optimales de C et ν par rapport au critère $CV(C, \nu)$ (les valeurs optimales étant $C = 1$ et $\nu = \frac{1}{10}$). Nous pouvons voir sur ces résultats que l'estimateur introduit par Ferraty et Vieu (2006) est légèrement meilleur. Comme nous le verrons plus loin, l'avantage de l'estimateur récursif sera sur le temps de calcul. Nous observons également comme attendu que l'erreur de prédiction diminue lorsque la taille de l'échantillon augmente ($n = 100$, $n = 200$, $n = 500$).

5.3.2 Choix de la semi-norme

Dans cette simulation, le paramètre ℓ est fixé égal à zéro et nous choisissons la fenêtre $h_i = \max_{j=1,\dots,n} \|X_j - \chi\| \, i^{-1/10}$ sur la base de ce que nous avons observé dans la simulation précédente. Nous considérons les semi-normes suivantes.

- La semi-norme $[PCA]$ est basée sur l'analyse en composantes principales des

	$n = 100$	$n = 200$	$n = 500$
Estimateur récursif	0.3022	0.2596	0.1993
	(0.6887)	(0.6275)	(0.5430)
Estimateur non récursif	0.2794	0.2143	0.1368
	(0.5512)	(0.5055)	(0.4208)

Table 5.1 Moyennes et écart-types de l'erreur quadratique de prédiction calculés sur 500 simulations pour l'estimateur non récursif et l'estimateur récursif.

courbes avec $q = 3$ composantes principales (voir Besse et al. (1997)), plus précisément

$$\|X_i - \chi\|_{PCA} = \sqrt{\sum_{j=1}^{q} \langle X_i - \chi, v_j \rangle^2},$$

où $\langle ., . \rangle$ est le produit scalaire usuel de l'espace des fonctions de carré intégrable et $(v_j)_{j \geq 1}$ est la suite des fonctions propres de l'opérateur de covariance empirique des courbes $X_1, \ldots, X_n$ (voir chapitre 2).

• La semi-norme $[FOU]$ est basée sur la décomposition des courbes dans la base usuelle de Fourier, avec $b = 8$ fonctions de base, plus précisément

$$\|X_i - \chi\|_{FOU} = \sqrt{\sum_{j=1}^{b} \left(a_{X_i,j} - a_{\chi,j}\right)^2},$$

où $(a_{X_i,j})_{j \geq 1}$ et $(a_{\chi,j})_{j \geq 1}$ sont les suites respectives des coefficients des courbes X_i et χ dans la base de Fourier.

• La semi-norme $[DERIV]$ est basée sur la décomposition des dérivées secondes des approximations des courbes par splines cubiques (avec $k = 8$ nœuds intérieurs pour les fonctions splines), plus précisément

$$\|X_i - \chi\|_{DERIV} = \sqrt{\langle \widetilde{X}_i - \widetilde{\chi}, \widetilde{X}_i - \widetilde{\chi} \rangle},$$

où $\widetilde{X}_i$ et $\widetilde{\chi}$ sont les approximations splines respectives des dérivées secondes des courbes X_i et χ.

Les résultats de cette simulations sont donnés dans la Table 5.2, toujours pour l'erreur quadratique de prédiction $MSPE$. Nous constatons sur cette simulation que la semi-norme $[PCA]$ montre de meilleurs résultats (ce qui semble naturel car pour un mouvement Brownien, l'approximation dans une base de Fourier où à l'aide de fonctions splines pour un processus irrégulier sera de mauvaise qualité). Cependant, il paraît clair qu'aucune semi-norme universelle qui surpasserait les autres ne semble

Semi-norme	$[PCA]$	$[FOU]$	$[DERIV]$
$MSPE$	0.3936	0.4506	0.4527
	(1.5190)	(1.5624)	(1.5616)

Table 5.2 Moyennes et écart-types de l'erreur quadratique de prédiction calculés sur 500 simulations pour différentes semi-normes.

exister, la performance d'une semi-norme va fortement être liée aux courbes X_i et à leur régularité. Ce fait est souligné par Ferraty et Vieu (2006).

5.3.3 Temps de calcul

Dans cette simulation, nous mettons en évidence l'avantage de l'estimateur récursif sur le non-récursif, vis-à-vis du temps de calcul nécessaire pour prédire une valeur de la variable réponse lorsque de nouvelles valeurs de la variable explicative arrivent séquentiellement dans la base de données. En effet, lorsqu'une nouvelle observation (X_{n+1}, Y_{n+1}) arrive dans l'échantillon, le calcul de l'estimateur récursif $\Psi_n^{[\ell]}$ requiert simplement une nouvelle itération de l'algorithme via sa valeur déjà calculée avec l'échantillon $(X_i, Y_i)_{i=1,\ldots,n}$, tandis que l'estimateur non-récursif doit être recalculé à partir de tout l'échantillon $(X_i, Y_i)_{i=1,\ldots,n+1}$. La simulation réalisée illustre la différence de temps de calcul entre les deux estimateurs dans ce genre de situation. À partir d'un échantillon initial $(X_i, Y_i)_{i=1,\ldots,n}$ de taille $n = 100$, nous considérons N observations additionnelles pour différentes valeurs de N. Nous comparons alors les temps de calcul cumulés pour obtenir l'estimateur récursif et l'estimateur non-récursif en ajoutant de nouvelles observations. Les calculs ont été réalisés sur une machine aux caractéristiques suivantes : CPU : Duo E4700 2.60 GHz, HD : 149 Go, Mémoire : 3.23 Go. Le nombre de nouvelles observations ajoutées à l'échantillon initial est $N \in \{1, 50, 100, 200, 500\}$. Nous choisissons comme semi-norme celle basée sur l'analyse en composantes principales des courbes avec 3 composantes principales, le paramètre ℓ est fixé égal à zéro, et nous choisissons la fenêtre $h_i = \max_{j=1,\ldots,n} \|X_j - \chi\| \, i^{-1/10}$.

Les temps de calcul (en secondes) sont donnés dans la Table 5.3. L'estimateur récursif montre un avantage très clair par rapport à l'estimateur non-récursif de Ferraty et Vieu (2006) vis-à-vis du temps de calcul dès lors qu'un nombre de plus en plus important de nouvelles observations arrive dans l'échantillon.

5.4 Application sur données réelles

Dans cette section, nous appliquons notre méthode d'estimation à un jeu de données réelles. Les données fonctionnelles sont particulièrement adaptées pour étudier les séries temporelles. En effet, une série temporelle sur une certaine période (par exemple, sur

	$N=1$	$N=50$	$N=100$	$N=200$	$N=500$
Estimateur récursif	0.125	0.484	0.859	1.563	3.656
Estimateur non récursif	0.047	1.922	5.594	21.938	152.719

Table 5.3 Temps de calcul cumulés (en secondes) pour le calcul de l'estimateur récursif et non-récursif lorsque N nouvelles observations arrivent dans l'échantillon pour différentes valeurs de N.

plusieurs années) peut être découpée en un certain nombre de courbes (par exemple, des courbes annuelles). Cette manière de traiter des séries temporelles comme des données fonctionnelles est inspirée de Bosq (2000). Nous considérons la série temporelle El Niño[1] qui donne la température mensuelle (en degrés Celsius) à la surface de l'eau au large du Pérou et de l'Équateur, de janvier 1982 à décembre 2011 (360 mois). Ces données sont tracées sur la Figure 5.1. De cette série, nous extrayons 30 courbes annuelles $X_1, \ldots, X_{30}$ de 1982 à 2011, discrétisées en $p = 12$ points. Ces courbes sont tracées sur la Figure 5.2. La variable d'intérêt le $j^{\text{ème}}$ mois de l'année i est la température le $j^{\text{ème}}$ mois de l'année $i+1$, en d'aitres termes, pour $j = 1, \ldots, 12$ et $i = 1, \ldots, 29$, on pose $Y_i^{[j]} = X_{i+1}(j)$.

Nous prédisons les valeurs $Y_{29}^{[1]}, \ldots, Y_{29}^{[12]}$, en d'autres termes, nous prédisons les valeurs de la courbe X_{30} chaque mois. L'estimateur récursif ainsi que l'estimateur non-récursif de Ferraty et Vieu (2006) sont calculés en choisissant comme paramètres (semi-norme, fenêtre et paramètre ℓ) comme dans les simulations. Nous analysons les résultats de prévisions en calculant le critère d'erreur quadratique moyenne de prévision

$$MSPE = \frac{1}{12}\sum_{j=1}^{12}\left(\widehat{Y}_{29}^{[j]} - Y_{29}^{[j]}\right)^2,$$

où $\widehat{Y}_{29}^{[j]}$ est calculé soit avec l'estimateur récursif ($MSPE = 0.5719$), soit avec l'estimateur non-récursif ($MSPE = 0.2823$). L'estimateur non-récursif de Ferraty et Vieu (2006) montre à nouveau son avantage par rapport à l'estimateur récursif en terme de précision d'erreur de prévision, tandis que l'estimateur récursif montre son avantage vis-à-vis du temps de calcul (0.128 s) par rapport à l'estimateur non-récursif (0.487 s), comme cela avait été constaté sur les simulations.

Bibliographie

I. Ahmad et P.E. Lin. Nonparametric sequential estimation of a multiple regression function. *Bulletin of Mathematical Statistics*, 17 :63–75, 1976.

[1] Disponible en ligne https ://www.math.univ-toulouse.fr/∼ferraty/SOFTWARES/NPFDA/

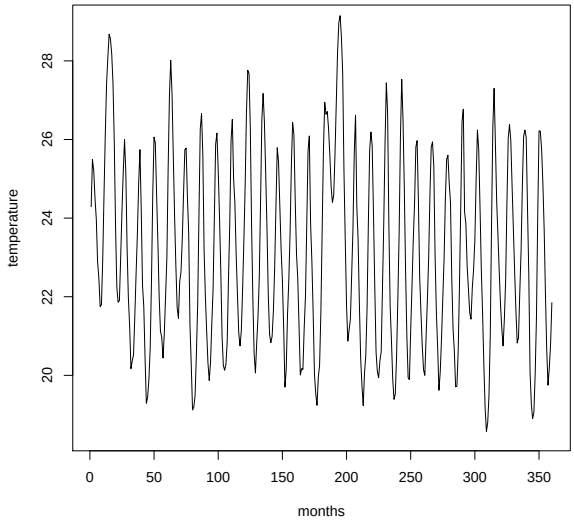

Figure 5.1 Série temporelle El Niño de janvier 1982 à décembre 2011 (données mensuelles).

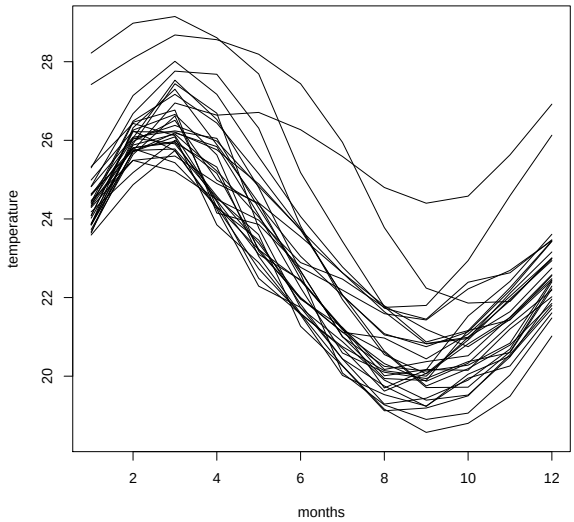

Figure 5.2 Courbes annuelles El Niño de 1982 à 2011 (données mensuelles).

A. Amiri. Recursive regression estimators with application to nonparametric prediction. *Journal of Nonparametric Statistics*, 24 :169–186, 2012.

A. Amiri, C. Crambes et B. Thiam. Recursive estimation of nonparametric regression with functional covariate. *Computational Statistics and Data Analysis*, 69 :154–172, 2014.

P. Besse, H. Cardot et F. Ferraty. Simultaneous nonparametric regression of unbalanced longitudinal data. *Computational Statistics and Data Analysis*, 24 :255–270, 1997.

D. Bosq. *Linear processes in function spaces.* Lecture Notes in Statistics, **149**, Springer, 2000.

D. Bosq et N. Cheze-Payaud. Optimal asymptotic quadratic error of nonparametric regression function estimates for a continuous-time process from sampled-data. *Statistics*, 32 :229–247, 1999.

L. Devroye et T.J. Wagner. Distribution-free consistency results in nonparametric discrimination and regression function estimation. *The Annals of Statistics*, 8 :231–239, 1980.

F. Ferraty et Y. Romain. *Handbook on functional data analysis and related fields.* Oxford University Press, Oxford, 2010.

F. Ferraty et P. Vieu. *Nonparametric functional data analysis :Theory and practice.* Springer-Verlag, New York, 2006.

F. Ferraty, A. Mas et P. Vieu. Nonparametric regression on functional data : Inference and practical aspects. *Australian and New Zealand Journal of Statistics*, 49 :267–286, 2007.

M. Rachdi et P. Vieu. Nonparametric regression for functional data : automatic smoothing parameter selection. *Journal of Statistical Planning and Inference*, 137 :2784–2801, 2007.

G. G. Roussas. Exact rates of almost sure convergence of a recursive kernel estimate of a probability densiy function : Application to regression and hazard rate estimation. *Journal of Nonparametric Statistics*, 1 :171–195, 1992.

Y. Slaoui. Recursive nonparametric regression estimation for independent functional data. *Statistica Sinica*, 2019.

Chapitre 6

Estimation de quantiles de régression par méthodes SVM

Contents

Les résultats présentés dans ce chapitre sont tirés de Crambes et al. (2011), Crambes et al. (2013) et Crambes et al. (2014).

6.1 Introduction

Dans ce chapitre, nous nous plaçons dans le cadre du modèle non-paramétrique fonctionnel (voir Ferraty et Vieu (2006)), écrit sous la forme

$$Y = \Psi(X) + \varepsilon, \tag{6.1}$$

où Y et ε sont à valeurs dans $\mathbb{R}$, X est à valeurs dans $L^2(I)$ et Ψ est un opérateur de $L^2(I)$ dans $\mathbb{R}$.

Chercher à modéliser la relation entre une variable réponse et une variable explicative est un problème central en statistiques. La littérature concernant la régression sur la moyenne est très vaste. Cependant, cette modélisation classique a certains désavantages : sensibilité aux valeurs aberrantes, inapproprié en présence de données multimodales ou asymétriques. Ces problèmes peuvent être contournés en utilisant un autre type de statistiques : les quantiles. En particulier, les quantiles conditionnels permettent de fournir une image plus complète de la distribution de Y sachant X pour différents niveaux de quantiles (voir Koenker (2005)). Habituellement, les quantiles conditionnels peuvent être déterminés de deux façons, soit en inversant une fonction de répartition conditionnelle, soit en résolvant un problème d'optimisation. Dans le cadre où la variable explicative est réelle ou multivariée, nous pouvons notamment citer les travaux concernant le modèle linéaire de quantiles conditionnels (voir Koenker et Bassett (1978)), ou encore les travaux dans un cadre non-paramétrique (voir notamment Chaudhuri (1991)). Même si l'approche non-paramétrique permet une estimation dans une classe très large, son inconvénient principal se situe dans un contexte de grande dimension, où le fléau de la dimension rend cette technique d'estimation inefficace. Pour contourner ce problème, plusieurs auteurs ont proposé des techniques de réduction de dimension, comme par exemple l'introduction de modèles additifs (voir Hastie et Tibshirani (1990)). Dans le contexte de la régression sur quantiles, on peut noter par exemple le travail de de Gooijer et Zerom (2003) où les auteurs considèrent estiment des quantiles conditionnels dans ce cadre multivarié en utilisant un estimateur à noyau de type Nadaraya-Watson de la fonction de répartition conditionnelle.

Nous souhaitons nous intéresser à l'estimation de quantiles de régression dans le cadre d'une variable explicative fonctionnelle. Le modèle (6.1) s'écrit

$$Y = \Psi_\tau(X) + \varepsilon_\tau, \tag{6.2}$$

où $\Psi = \Psi_\tau$ est l'opérateur quantile à estimer, sur la base d'un échantillon $(X_i, Y_i)_{i=1,\dots,n}$, et le bruit $\varepsilon = \varepsilon_\tau$ vérifie $\mathbb{P}(\varepsilon_\tau \leq 0|X) = \tau$ où $\tau \in]0;1[$ est l'ordre du quantile. Dans ce cadre d'estimation de quantiles conditionnels lorsque la variable explicative est fonctionnelle, les principaux travaux sont ceux de Cardot et al. (2005) (l'estimateur est basé sur des fonctions splines), Ferraty et al. (2005) (l'estimateur est de type à noyau), Chen et Müller (2012) (les courbes correspondant à la variable explicative sont décomposées dans la base des fonctions propres de l'opérateur de covariance). Dans le cas où nous disposons de plusieurs variables explicatives fonctionnelles, nous notons à ce jour peu de travaux, réalisés pour l'estimation de la moyenne conditionnelle (voir Müller et Yao (2008), Ferraty et al. (2013)).

Dans notre travail, nous considérons un modèle additif dérivé de (6.2), et nous introduisons un estimateur basé sur la méthode Support Vector Machine (SVM), basé

notamment sur les travaux de Vapnik (1998), Schölkopf et Smola (2002), Steinwart et Christmann (2008) dans le cas où la variable explicative est réelle ou multivariée. Cette méthode d'apprentissage est basée sur la minimisation d'un risque pénalisé dans un espace de Hilbert à noyau reproduisant (RKHS, Reproducing Kernel Hilbert Space, voir Berlinet et Thomas-Agnan (2004)). Lorsque la variable explicative fonctionnelle, les travaux utilisant la méthode SVM sont peu nombreux. On peut citer Rossi et Villa-Vialaneix (2006) et Gonzalez et Munoz (2010) dans le cadre de la classification, et Preda (2007) dans le cadre de la régression sur la moyenne.

6.2 Estimateur

Nous estimons le quantile conditionnel d'ordre τ en utilisant le fait que le quantile conditionnel est solution du problème de minimisation suivant

$$\Psi_\tau(x) = \min_{a\in\mathbb{R}} \mathbb{E}\left(\rho_\tau\left(Y-a\right)|X=x\right), \tag{6.3}$$

pour tout $x \in L^2(I)$, où la fonction ρ_τ est la "check function" définie par $\rho_\tau(u) = |u| + (2\tau-1)u$ pour tout $u \in \mathbb{R}$. Pour estimer Ψ_τ, l'idée est d'utiliser le problème de minimisation (6.3). Cependant, dans ce contexte où la variable explicative X est fonctionnelle, nous sommes confrontés à un problème mal posé lié à la non-existence d'un inverse borné de l'opérateur de covariance de X. L'une des façons usuelles de contourner ce problème est de régulariser le problème d'optimisation en pénalisant d'une certaine façon la fonction cible Ψ_τ. Les techniques SVM sont une classe d'algorithmes d'apprentissage permettant de déterminer une solution d'un problème de minimisation convexe pénalisé dans un espace de Hilbert à noyau reproduisant. Les étapes de la procédure d'estimation sont détaillées dans les paragraphes suivants.

6.2.1 Projection

Nous commençons par projeter les courbes $X_1, \ldots, X_n$ sur une base orthonormée de fonctions $(\phi_j)_{j\geq 1}$. Pour tout $i = 1, \ldots, n$, la fonction X_i va se décomposer sous la forme

$$X_i(t) = \sum_{j=1}^{+\infty} X_{ij}\phi_j(t),$$

pour tout $t \in I$, les coefficients de la décomposition étant donnés par

$$X_{ij} = \int_I X_i(t)\phi_j(t)\,dt.$$

Le fait de commencer par projeter les courbes $X_1, \ldots, X_n$ dans une base est une première étape pour réduire la dimension du problème initial. Cette méthodologie est fréquemment utilisée, voir par exemple Biau et al. (2005). Cela permet de résumer l'information

présente dans les courbes dans un vecteur de dimension en général petite. On choisit d'utiliser un nombre d de coefficients de base, et ainsi d'approximer les courbes par

$$\sum_{j=1}^{d} X_{ij}\phi_j(t).$$

Dans la suite, on note $\mathbf{X}_i^{(d)}$ le vecteur de dimension d constitué des d premiers coefficients de la décomposition de la fonction X_i dans la base $(\phi_j)_{j\geq 1}$.

6.2.2 Résolution du problème de minimisation

Soit $\mathcal{K} : L^2(I) \times L^2(I) \to \mathbb{R}$ un noyau (i.e. une fonction symétrique définie positive), d'espace de Hilbert à noyau reproduisant associé $\mathcal{H}_\mathcal{K}$. On note $\mathcal{K}^{(d)}$ la restriction de $\mathcal{K}$ à l'espace engendré par les d premières fonctions de base $\phi_1, \ldots, \phi_d$, assimilé à $\mathbb{R}^d$. À partir de maintenant, l'échantillon est séparé en deux : un échantillon d'apprentissage $(X_i, Y_i)_{i=1,\ldots,\ell}$ et un échantillon de test $(X_i, Y_i)_{i=\ell+1,\ldots,n}$. Le théorème de représentation (voir Kimeldorf et Wahba (1971)) permet de chercher une solution au problème de minimisation

$$\min_{\Psi\in\mathcal{H}_\mathcal{K}} \left\{ \frac{1}{\ell}\sum_{i=1}^{\ell} \rho_\tau \left(Y_i - \Psi(X_i)\right) + \lambda \left\|\Psi\right\|^2_{\mathcal{H}_\mathcal{K}} \right\},$$

où $\lambda > 0$ est un paramètre de régularisation qui vise à réduire l'effet de sur-apprentissage. La solution à ce problème de minimisation aura alors une représentation de la forme

$$\sum_{i=1}^{\ell} \alpha_i \mathcal{K}(., X_i),$$

où $(\alpha_i)_{i=1,\ldots,\ell} \in \mathbb{R}^\ell$ (voir Schölkopf et Smola (2002)). En adaptant cette méthodologie à notre cadre, nous allons chercher $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_i)_{i=1,\ldots,\ell} \in \mathbb{R}^\ell$ solution du problème de minimisation

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^\ell} \left\{ \frac{1}{\ell}\sum_{i=1}^{\ell} \rho_\tau \left(Y_i - \sum_{j=1}^{\ell} \alpha_j \mathcal{K}^{(d)}(X_i^{(d)}, X_j^{(d)}) \right) + \lambda \boldsymbol{\alpha}' \mathbf{K}^{(d)} \boldsymbol{\alpha} \right\}, \tag{6.4}$$

où la matrice $\mathbf{K}^{(d)} = \left(\mathcal{K}^{(d)}(X_i^{(d)}, X_j^{(d)})\right)_{i,j=1,\ldots,\ell}$ est la matrice de Gram associée au noyau $\mathcal{K}^{(d)}$.

6.2.3 Sélection des paramètres

L'estimateur défini dans le paragraphe précédent dépend de plusieurs paramètres : la dimension d correspondant au nombre de fonctions de base, le noyau $\mathcal{K}$ et le paramètre de régularisation λ. Nous allons utiliser une procédure de sélection de ces paramètres de façon analogue à Biau et al. (2005). On se donne un ensemble $\mathcal{V}$ de triplés $\mathbf{V} = (d, \mathcal{K}, \lambda)$ et on cherche le triplé $\mathbf{V}^\star$ qui minimise l'erreur pénalisée suivante, calculée sur l'échantillon de test

$$\min_{\mathbf{V}\in\mathcal{V}} \left\{ \frac{1}{n-\ell} \sum_{i=\ell+1}^{n} \rho_\tau \left(Y_i - \sum_{j=1}^{\ell} \widehat{\alpha}_j \mathcal{K}^{(d)}(X_i^{(d)}, X_j^{(d)}) \right) + \frac{\beta^{(d)}}{\sqrt{n-\ell}} \right\}, \tag{6.5}$$

où les coefficients $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_\ell$ sont estimés en résolvant le problème de minimisation (6.4) avec les paramètres $(d, \mathcal{K}, \lambda)$ choisis dans $\mathcal{V}$, et le terme de pénalisation avec le paramètre $\beta^{(d)}$ permet d'éviter le sur-apprentissage. L'estimateur final de Ψ_τ est l'estimateur $\widehat{\Psi}_n$ solution du problème de minimisation (6.4) avec les paramètres $\mathbf{V}^\star = (d^\star, \mathcal{K}^\star, \lambda^\star)$ solutions du problème de minimisation (6.5). Notons pour finir que nous incluons nécessairement le noyau Gaussien parmi l'ensemble des noyaux possibles, noté $\mathcal{Z}$, de façon à avoir un noyau universel parmi les noyaux possibles. Un noyau est universel (voir Steinwart (2001)) lorsque le RKHS qui lui est associé est dense dans l'ensemble des fonctions continues.

6.3 Modèle additif

On se place dans un cadre plus général que la section précédente, en considérant que l'on dispose de s variables explicatives : $X^1, \ldots, X^s$. Le modèle considéré est un modèle additif

$$Y = \sum_{r=1}^{s} \Psi_\tau^r(X^r) + \varepsilon_\tau, \tag{6.6}$$

où $\Psi_\tau^1, \ldots, \Psi_\tau^s$ sont les opérateurs quantile à estimer, sur la base d'un échantillon $(X_i, Y_i)_{i=1,\ldots,n}$, et le bruit ε_τ vérifie $\mathbb{P}(\varepsilon_\tau \leq 0 | X^1, \ldots, X^s) = \tau$ où $\tau \in]0;1[$ est l'ordre du quantile. La procédure d'estimation introduite dans la section précédente se généralise naturellement à ce cadre.

6.3.1 Projection

Nous commençons par projeter les courbes X_i^r pour $r = 1, \ldots, s$ et $i = 1, \ldots, n$ sur une base orthonormée de fonctions $(\phi_j)_{j\geq 1}$. On choisit des dimensions $d^1, \ldots, d^s$ de projection, pour approximer les courbes par

$$\sum_{j=1}^{d^r} X_{ij}^r \phi_j(t).$$

Dans la suite, on note $\mathbf{X}_i^{r,(d^r)}$ le vecteur de dimension d^r constitué des d^r premiers coefficients de la décomposition de la fonction X_i^r dans la base $(\phi_j)_{j\geq 1}$.

6.3.2 Résolution du problème de minimisation

Soit $\mathcal{K}^r : L^2(I)\times L^2(I) \to \mathbb{R}$ un noyau, d'espace de Hilbert à noyau reproduisant associé $\mathcal{H}_{\mathcal{K}^r}$. On note $\mathcal{K}^{(d^r)}$ la restriction de $\mathcal{K}^r$ à l'espace engendré par les d^r premières fonctions de base $\phi_1,\ldots,\phi_d$, assimilé à $\mathbb{R}^{d^r}$. Comme dans la section précédente, l'échantillon est séparé en deux : un échantillon d'apprentissage $(X_i^r, Y_i)_{r=1,\ldots,s, i=1,\ldots,\ell}$ et un échantillon de test $(X_i^r, Y_i)_{r=1,\ldots,s, i=\ell+1,\ldots,n}$. Nous allons chercher $\widehat{\boldsymbol{\alpha}}^{\boldsymbol{r}} = (\widehat{\alpha}_i^r)_{i=1,\ldots,\ell} \in \mathbb{R}^\ell$ pour $r = 1,\ldots,s$ solution du problème de minimisation

$$\min_{\boldsymbol{\alpha}^1,\ldots,\boldsymbol{\alpha}^s\in\mathbb{R}^\ell} \left\{ \frac{1}{\ell}\sum_{i=1}^{\ell} \rho_\tau\left(Y_i - \sum_{r=1}^{s} [\mathbf{K}]_{i.}^{r,d^r}\,\boldsymbol{\alpha}^r\right) + \sum_{r=1}^{s} \lambda^r (\boldsymbol{\alpha}^r)' \mathbf{K}^{r,(d^r)} \boldsymbol{\alpha}^r \right\}, \tag{6.7}$$

où la matrice $\mathbf{K}^{r,(d^r)} = \left(\mathcal{K}^{r,(d^r)}(X_i^{r,(d^r)}, X_j^{r,(d^r)})\right)_{i,j=1,\ldots,\ell}$ est la matrice de Gram associée au noyau $\mathcal{K}^{r,(d^r)}$ et les paramètres $\lambda^1,\ldots,\lambda^s$ sont des paramètres de régularisation.

6.3.3 Sélection des paramètres

Il nous faut à nouveau sélectionner plusieurs paramètres : les dimensions $d^1,\ldots,d^s$ correspondant au nombre de fonctions de base choisies pour projeter chacune des variables, les noyaux $\mathcal{K}^1,\ldots,\mathcal{K}^s$ et les paramètres de régularisation $\lambda^1,\ldots,\lambda^s$. Nous utilisons la même procédure de sélection de ces paramètres que dans la section précédente. On se donne un ensemble $\mathcal{V}$ de $3s$-uplés $\mathbf{V} = (d^1,\ldots,d^s,\mathcal{K}^1,\ldots,\mathcal{K}^s,\lambda^1,\ldots,\lambda^s)$ et on cherche $\mathbf{V}^\star$ qui minimise l'erreur pénalisée suivante, calculée sur l'échantillon de test

$$\min_{\mathbf{V}\in\mathcal{V}} \left\{ \frac{1}{n-\ell}\sum_{i=\ell+1}^{n} \rho_\tau\left(Y_i - \sum_{r=1}^{s}\sum_{j=1}^{\ell} \widehat{\alpha}_j^r \mathcal{K}^{r,(d^r)}(X_i^{r,(d^r)}, X_j^{r,(d^r)})\right) + \max_{r=1,\ldots,s} \frac{\beta^{(d^r)}}{\sqrt{n-\ell}} \right\}, \tag{6.8}$$

où les coefficients $\widehat{\alpha}_j^r$ sont estimés en résolvant le problème de minimisation (6.7) avec les paramètres $(d^1,\ldots,d^s,\mathcal{K}^1,\ldots,\mathcal{K}^s,\lambda^1,\ldots,\lambda^s)$ choisis dans $\mathcal{V}$. L'estimateur final de $\Psi_\tau^1,\ldots,\Psi_\tau^s$ est l'estimateur $\widehat{\Psi}_n^1,\ldots,\widehat{\Psi}_n^1$ solution du problème de minimisation (6.7) avec les paramètres $\mathbf{V}^\star$ solutions du problème de minimisation (6.8).

6.4 Résultats asymptotiques

Dans cette section, nous présentons les principaux résultats de convergence obtenus sur les estimateurs présentés dans les paragraphes précédents. Les preuves de ces résultats sont données dans les articles Crambes et al. (2011), Crambes et al. (2013) et Crambes et al. (2014). Les points-clés des preuves sont basés sur des applications d'inégalités exponentielles et l'utilisation de nombres de couverture (voir Zhou (2002)).

6.4.1 Consistence

Les hypothèses permettant d'obtenir un résultat de consistence de l'estimateur sont les suivantes.

(H.1) Nous supposons que l'opérateur Ψ_τ est γ-Höldérien avec $\gamma > 0$: il existe une constante $C_1 > 0$ telle que pour tous $a, b \in L^2(I)$

$$|\Psi_\tau(a) - \Psi_\tau(b)| \leq C_1 \|a - b\|_\infty^\gamma .$$

(H.2) Nous supposons que X prend p.s. ses valeurs dans un sous-espace borné de $L^2(I)$ et que Y est p.s. bornée.

(H.3) Nous supposons qu'il existe un réel $R > 0$ tel que, pour tout $d \geq 1$

$$\left\| \Psi_\tau^{(d)} \right\|_{\mathcal{H}_\mathcal{K}} \leq R,$$

et qu'il existe un réel $M > 0$ tel que, pour tout $d \geq 1$

$$\left| \rho_\tau \left(Y - \Psi_\tau^{(d)}(X^{(d)}) \right) | X^{(d)} = x^{(d)} \right| \leq M.$$

(H.4) Nous supposons que

$$\lim_{d \to +\infty} \beta^{(d)} = +\infty,$$
$$\lim_{\substack{d \to +\infty \\ n-\ell \to +\infty}} \frac{\beta^{(d)}}{\sqrt{n-\ell}} = 0.$$

(H.5) Soit $C_{\mathcal{K}^{(d)}}$ qui ne dépend que du noyau $\mathcal{K}$ et de la dimension d tel que

$$\lim_{\substack{d\to+\infty\\ n-\ell\to+\infty}} \frac{C_{\mathcal{K}^{(d)}}}{\sqrt{n-\ell}} = 0,$$

$$\Delta := \sum_{d\geq 1} |\mathcal{Z}| \exp\left\{ C_{\mathcal{K}^{(d)}} \left(\ln \frac{8R}{\frac{\beta^{(d)}}{\sqrt{n-\ell}}} \right)^{d+1} \right\} \exp\left(\frac{-\left(\beta^{(d)}\right)^2}{8M^2} \right) < +\infty.$$

Nous sommes maintenant en mesure d'énoncer le résultat de consistence.

Theorème 6.1. *Sous les hypothèses (H.1)-(H.5), l'estimateur $\widehat{\Psi}_n$ vérifie l'inégalité oracle suivante, pour $x \in L^2(I)$*

$$\begin{aligned}
&\mathbb{E}\left[\rho_\tau\left(\widehat{\Psi}_n(x) - \Psi_\tau(x)\right)\right] \\
\leq\ &\inf_{d\geq 1}\left\{ \mathbb{E}\left[\rho_\tau(Y - \Psi_\tau^{(d)}(X^{(d)}))|X^{(d)} = x^{(d)}\right] - \mathbb{E}\left[\rho_\tau(Y - \Psi_\tau(X))|X = x\right] \right. \\
&\left. + \inf_{\lambda,\mathcal{K}} \mathbb{E}\left[\rho_\tau\left(\widehat{\Psi}_{d,\lambda,\mathcal{K}}(x^{(d)}) - \widehat{\Psi}_\tau^{(d)}(x^{(d)})\right)\right] + \frac{\beta^{(d)}}{\sqrt{n-\ell}} \right\} \\
&+ M\sqrt{\frac{8ln(\Delta)}{n-\ell}} + \frac{\sqrt{8}M}{\sqrt{(n-\ell)ln(\Delta)}}.
\end{aligned}$$

Le premier terme $\mathbb{E}\left[\rho_\tau(Y - \Psi_\tau^{(d)}(X^{(d)}))|X^{(d)} = x^{(d)}\right] - \mathbb{E}\left[\rho_\tau(Y - \Psi_\tau(X))|X = x\right]$ dans l'inégalité oracle précédente est le prix à payer pour utiliser une approximation finie-dimensionnelle des observations. De plus, à une dimension d fixée, le second terme $\mathbb{E}\left[\rho_\tau\left(\widehat{\Psi}_{d,\lambda,\mathcal{K}}(x^{(d)}) - \widehat{\Psi}_\tau^{(d)}(x^{(d)})\right)\right]$ va lui aussi tendre vers zéro en utilisant les résultats de Steinwart et Christmann (2008) en dimension finie. Ainsi, le résultat précédent a le corollaire suivant.

Corollaire 6.2. *Sous les hypothèses du Théorème 6.1, nous avons*

$$\lim_{\substack{n\to+\infty\\ n-\ell\to+\infty}} \mathbb{E}\left[\rho_\tau\left(\widehat{\Psi}_n(x) - \Psi_\tau(x)\right)\right] = 0.$$

Pour conclure ce paragraphe, nous ajoutons qu'un résultat analogue a été obtenu pour le modèle additif (voir Crambes et al. (2014)) au prix d'hypothèses similaires adaptées au cadre additif.

6.4.2 Vitesse de convergence

Nous pouvons obtenir une vitesse de convergence de l'estimateur $\widehat{\Psi}_n$. Nous considérons pour cela les hypothèses (H.1)-(H.5), auxquelles nous rajoutons les hypothèses suivantes.

(H.6) Nous supposons que la fonction X est δ-Höldérienne avec $\delta > 0$: il existe une constante $C_2 > 0$ telle que pour tous $t, s \in I$

$$|X(t) - X(s)| \leq C_2 |t - s|^{\delta}.$$

(H.7) Nous supposons qu'il existe une constante $C_3 > 0$ et un réel $\alpha \in]0; 1]$ tels que

$$\mathbb{E}\left[e_\tau\left(X^{(d)}, Y\right)\right]^{\alpha} \geq C_3 \mathbb{E}\left|\Psi^{(d)}(X^{(d)}) - \Psi_\tau^{(d)}(X^{(d)})\right|,$$

pour tout $\Psi^{(d)}$ tel que $\left\|\Psi^{(d)}\right\|_{\mathcal{H}_{\mathcal{K}^{(d)}}} < +\infty$, où

$$\begin{aligned} e_\tau\left(X^{(d)}, Y\right) &= \left(\Psi^{(d)}(X^{(d)}) - Y\right) \mathbb{1}_{\left\{\Psi_\tau^{(d)}(X^{(d)}) \leq Y \leq \Psi^{(d)}(X^{(d)})\right\}} \\ &\quad + \left(Y - \Psi^{(d)}(X^{(d)})\right) \mathbb{1}_{\left\{\Psi^{(d)}(X^{(d)}) \leq Y \leq \Psi_\tau^{(d)}(X^{(d)})\right\}}. \end{aligned}$$

Nous sommes maintenant en mesure d'énoncer le résultat donnant une vitesse de convergence de l'estimateur.

Theorème 6.3. *Sous les hypothèses (H.1)-(H.7), en considérant une suite $(v_n)_{n\geq 1}$ vérifiant, lorsque n et $n - \ell$ tendent vers $+\infty$*

$$\begin{aligned} & v_n \sqrt{n - \ell} \to +\infty, \\ & d^{\gamma\delta} v_n \to +\infty, \\ & \frac{\beta^{(d)}}{v_n \sqrt{n - \ell}} \to 0, \\ & v_n^{-1} \left(\frac{(ln\, n)^2}{n}\right)^{1/(2-\alpha)} \to 0, \end{aligned}$$

alors l'estimateur $\widehat{\Psi}_n$ vérifie, pour $x \in L^2(I)$

$$\widehat{\Psi}_n(x) - \Psi_\tau(x) = O_{\mathbb{P}}(v_n).$$

Ce résultat a la conséquence immédiate suivante, permettant d'expliciter une vitesse.

Corollaire 6.4. *Sous les hypothèses du Théorème 6.3, en prenant par exemple* $v_n = \left(\frac{(ln\, n)^3}{n}\right)^{1/(2-\alpha)}$, *nous avons*

$$\widehat{\Psi}_n(x) - \Psi_\tau(x) = O_{\mathbb{P}}\left(\left(\frac{(ln\, n)^3}{n}\right)^{1/(2-\alpha)}\right).$$

Le résultat précédent fournit une vitesse de convergence en probabilité pour notre estimateur, dépendant du paramètre α. Lorsque α s'approche de 1, la vitesse est quasiment paramétrique, à un terme logarithmique près. Lorsque α s'approche de zéro, la vitesse se dégrade. Un exemple est donné dans Li et al. (2007).

6.5 Applications

6.5.1 Implémentation numérique

Le problème d'optimisation (6.4) ne se solutionne pas de façon immédiate, en raison de la non-dérivabilité en zéro de la fonction ρ_τ. En particulier, il n'est pas possible d'obtenir une forme explicite d'une solution de ce problème d'optimisation. Nous contournons cette difficulté en adoptant l'algorithme des moindres carrés itérés pondérés, applicable dans ce cas (voir Ruppert et Caroll (1988)). La description de cet algorithme dans notre cadre est décrite ci-dessous. Cet algorithme a permis l'implémentation de l'estimateur $\widehat{\Psi}_n$. Nous avons également étendu cette implémentation au cadre du modèle additif (voir Crambes et al. (2014)). Le principe est le même que pour le modèle avec une seule variable explicative, en couplant l'algorithme des moindres carrés itérés pondérés avec un algorithme backfitting. Les résultats obtenus en simulations, disponibles dans les articles Crambes et al. (2013) et Crambes et al. (2014), montrent la consistence de la méthode d'estimation, avec une comparaison favorable par rapport à d'autres méthodes d'estimation de quantiles de régression de la littérature (Cardot et al. (2005), Ferraty et al. (2005)).

- Notations préliminaires : soit $\boldsymbol{\alpha} \in \mathbb{R}^\ell$, nous considérons la fonction Γ_i définie pour tout $i = 1, \ldots, \ell$ par

$$\Gamma_i(\tau) = \begin{cases} 2\tau \text{ si } Y_i - [\mathbf{K}]_{i.}^{(d)}\,\boldsymbol{\alpha} \geq 0, \\ 2(1-\tau) \text{ si } Y_i - [\mathbf{K}]_{i.}^{(d)}\,\boldsymbol{\alpha} < 0, \end{cases}$$

et le problème de minimisation (6.4) s'écrit sous la forme

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^\ell}\left\{\frac{1}{2\lambda\ell}\sum_{i=1}^{\ell}\Gamma_i(\tau)\left|Y_i - [\mathbf{K}]_{i.}^{(d)}\,\boldsymbol{\alpha}\right| + \frac{1}{2}\boldsymbol{\alpha}'\mathbf{K}^{(d)}\boldsymbol{\alpha}\right\}. \tag{6.9}$$

- Initialisation de l'algorithme : nous déterminons $\widehat{\boldsymbol{\alpha}}^{(1)}$ solution du problème d'optimisation quadratique

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^\ell}\left\{\frac{1}{2\lambda\ell}\sum_{i=1}^{\ell}\Gamma_i(\tau)\left(Y_i-[\mathbf{K}]_{i.}^{(d)}\boldsymbol{\alpha}\right)^2+\frac{1}{2}\boldsymbol{\alpha}'\mathbf{K}^{(d)}\boldsymbol{\alpha}\right\},$$

qui donne comme solution

$$\widehat{\boldsymbol{\alpha}}^{(1)}=\left(\lambda\ell\mathbf{I}+\mathbf{K}^{(d)}\right)^{-1}\mathbf{Y},$$

où $\mathbf{I}$ est la matrice identité de taille ℓ et $\mathbf{Y}=(Y_1,\ldots,Y_\ell)'$.

- Récurrence : connaissant $\widehat{\boldsymbol{\alpha}}^{(z)}$, nous déterminons $\widehat{\boldsymbol{\alpha}}^{(z+1)}$ solution du problème d'optimisation

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^\ell}\left\{\frac{1}{2\lambda\ell}\sum_{i=1}^{\ell}\frac{\Gamma_i^z(\tau)\left(Y_i-[\mathbf{K}]_{i.}^{(d)}\boldsymbol{\alpha}\right)^2}{\left[\left(Y_i-[\mathbf{K}]_{i.}^{(d)}\boldsymbol{\alpha}\right)^2+\eta 2\right]^{1/2}}+\frac{1}{2}\boldsymbol{\alpha}'\mathbf{K}^{(d)}\boldsymbol{\alpha}\right\},$$

où η^2 est une petite quantité strictement positive permettant d'éviter des problèmes numériques de dénominateurs trop proches de zéro et $\Gamma_i^z(\tau)$ est $\Gamma_i(\tau)$ à l'étape z de l'algorithme, c'est-à-dire calculé avec $\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}^{(z)}$. La stratégie de l'algorithme des moindres carrés itérés pondérés consiste ainsi à remplacer la valeur absolue par un terme au carré divisé par la racine carrée de celui-ci. En définissant la matrice $\mathbf{M}^z$ comme la matrice diagonale de taille ℓ d'éléments diagonaux donnés pour $i=1,\ldots,\ell$ par

$$\frac{\Gamma_i^z(\tau)}{\left[\left(Y_i-[\mathbf{K}]_{i.}^{(d)}\boldsymbol{\alpha}\right)^2+\eta 2\right]^{1/2}},$$

on obtient

$$\widehat{\boldsymbol{\alpha}}^{(z+1)}=\left(\lambda\ell\mathbf{I}+\mathbf{M}^z\mathbf{K}^{(d)}\right)^{-1}\mathbf{M}^z\mathbf{Y}.$$

- Critère d'arrêt : on choisit d'arrêter l'algorithme lorsque $\left\|\widehat{\boldsymbol{\alpha}}^{(z+1)}-\widehat{\boldsymbol{\alpha}}^{(z)}\right\|<err$, où err est une constante fixée par l'utilisateur.

6.5.2 Application à la prévision de pics de pollution à l'ozone

6.5.2.1 Description des données

Nous avons appliqué notre méthode d'estimation de quantiles de régression sur des données de pollution de l'air. Les données concernent des mesures de la qualité de l'air autour de la ville de Toulouse durant les périodes chaudes (du 15 mai au 15 septembre) des années 1997 à 2000. L'échantillon est de 474 jours (22 jours sont manquants pour des raisons de pannes techniques sur les appareils de mesure). Ces données sont présentées plus précisément et traitées par d'autres méthodes dans les travaux Aneiros-Perez et al. (2004), Cardot et al. (2007), Ferraty et Vieu (2009). Nous disposons de cinq variables explicatives fonctionnelles X^1, X^2, X^3, X^4, X^5, qui sont des courbes journalières mesurées heure par heure. Trois de ces variables sont des composants chimiques de l'air : concentration en ozone (O_3), concentration en monoxide d'azote (NO), concentration en dioxyde d'azote (NO_2). Les deux autres variables sont climatologiques : la direction du vent (WD) et la vitesse du vent (WS). La variable réponse scalaire Y est définie par $Y_i = \max_{t\in[0,24]} O_{3,i}(t)$ pour tout $i = 1, \ldots, 474$, ce qui correspond au maximum de concentration d'ozone le jour i. L'objectif est de prédire le pic de pollution un jour donné en connaissant les courbes journalières des cinq variables explicatives la veille de ce jour. Afin de nous comparer à la méthode de Cardot et al. (2007), l'échantillon de 474 jours a été partagé aléatoirement en un échantillon d'apprentissage (de taille $\ell = 332$) et un échantillon de test (de taille $m = 142$). L'échantillon d'apprentissage est utilisé afin de sélectionner les paramètres et estimer, et l'échantillon de test est utilisé pour comparer les prédictions. Les comparaisons se font sur la base de critères présentés dans le paragraphe suivant.

6.5.2.2 Critères de comparaison

Nous considérons les trois critères suivants :

- l'erreur moyenne absolue

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|Y_i - \widehat{Y}_i\right|,$$

- l'erreur quadratique relative (où $\overline{Y}_\ell$ désigne la moyenne empirique de l'échantillon d'apprentissage $(Y_i)_{i=1,\ldots,\ell}$)

$$QRE = \frac{\sum_{i=1}^{m}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{m}\left(Y_i - \overline{Y}_\ell\right)^2},$$

	Méthode			
	Variable explicative O_3		Variables explicatives O_3, NO, NO_2, WD, WS	
	Approche SVM (noyau Gaussien)	Approche Cardot et al. (2007)	Approche SVM (noyau Gaussien)	Approche Cardot et al. (2007)
MAE	12.154	12.332	11.643	11.864
QRE	0.415	0.425	0.383	0.397
qRE	0.651	0.661	0.623	0.638

Table 6.1 Comparaison des erreurs de prévision pour différentes méthodes de prévision.

- l'erreur relative sur quantile (où $q_\tau(Y)_\ell$ désigne le quantile empirique d'ordre τ de l'échantillon d'apprentissage $(Y_i)_{i=1,\ldots,\ell}$)

$$qRE = \frac{\sum_{i=1}^{m} \rho_\tau \left(Y_i - \widehat{Y}_i\right)}{\sum_{i=1}^{m} \rho_\tau \left(Y_i - q_\tau(Y)_\ell\right)}.$$

6.5.2.3 Résultats

Nous avons construit plusieurs modèles additifs avec les différentes variables explicatives afin de trouver la meilleure erreur de prévision sur l'échantillon de test. La prédiction a été réalisée à l'aide de la médiane conditionnelle, correspondant à un ordre $\tau = 0.5$ de quantile. Les résultats, présentés dans Crambes et al. (2014), montrent que la variable la plus influente est la concentration en ozone, tandis que les quatre autres variables explicatives ont un pouvoir prédictif beaucoup plus faible, la prédiction se retrouve très légèrement améliorée lorsque ces variables sont ajoutées au modèle. Les meilleurs résultats sont obtenus avec un noyau Gaussien, sans projection préalable des courbes. Notre étude améliore les résultats en comparaison avec ceux obtenus dans Cardot et al. (2007), comme cela apparaît dans la Table 6.1.

Bibliographie

G. Aneiros-Perez, H. Cardot, G.E. Perez et P. Vieu. Maximum ozone concentration forecasting by functional nonparametric approaches. *Environmetrics*, 15 :675–685, 2004.

A. Berlinet et C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publisher, 2004.

G. Biau, F. Bunea et M.H. Wegkamp. Functional classification in hilbert spaces. *IEEE Transactions on Information Theory*, 51 :2163–2172, 2005.

H. Cardot, C. Crambes et P. Sarda. Quantile regression when the covariates are functions. *Journal of Nonparametric Statistics*, 17 :841–856, 2005.

H. Cardot, C. Crambes et P. Sarda. Ozone pollution forecasting using conditional mean and conditional quantiles with functional covariates. *Statistical methods for biostatistics and related fields, eds W. Härdle, Y. Mori, P. Vieu*, pages 509–529, 2007.

P. Chaudhuri. Global nonparametric estimation of conditional quantile functions and their derivatives. *Journal of Multivariate Analysis*, 39 :246–269, 1991.

K. Chen et H.-G. Müller. Conditional quantile analysis when covariates are functions, with application to growth data. *Journal of the Royal Statistical Society, Series B*, 74 :67–89, 2012.

C. Crambes, A. Gannoun et Y. Henchiri. Weak consistency of the support vector machine quantile regression approach when covariates are functions. *Statistics and Probability Letters*, 81(12) :1847–1858, 2011.

C. Crambes, A. Gannoun et Y. Henchiri. Support vector machine quantile regression approach for functional data : Simulation and application studies. *Journal of Multivariate Analysis*, 121 :50–68, 2013.

C. Crambes, A. Gannoun et Y. Henchiri. Modelling functional additive quantile regression using support vector machine approach. *Journal of Nonparametric Statistics*, 26 :639–668, 2014.

J. de Gooijer et D. Zerom. On additive conditional quantiles with high dimensional covariates. *Journal of the American Statistical Association*, 98 :135–146, 2003.

F. Ferraty et P. Vieu. *Nonparametric functional data analysis : Theory and practice.* Springer-Verlag, New York, 2006.

F. Ferraty et P. Vieu. Additive prediction and boosting for functional data. *Computational Statistics and Data Analysis*, 53 :1400–1413, 2009.

F. Ferraty, A. Rabhi et P. Vieu. Conditional quantiles for dependent functional data with application to the climatic El Niño phenomenon. *Sankhya, Special Issue on Quantile Regression and Related Methods*, 67 :378–398, 2005.

F. Ferraty, A. Goia, E. Salinelli et P. Vieu. Functional projection pursuit regression. *Test*, 22 :293–320, 2013.

J. Gonzalez et A. Munoz. Representing functional data in reproducing kernel hilbert spaces with applications to clustering and classification. *Statistics and Econometrics*, 13 :10–27, 2010.

T. Hastie et R. Tibshirani. *Generalized additive models.* Chapman and Hall, London, 1990.

G.S. Kimeldorf et G. Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1) :82–95, 1971.

R. Koenker. *Quantile Regression.* Cambridge University Press, 2005.

R. Koenker et G. Bassett. Regression quantiles. *Econometrica*, 46 :33–50, 1978.

Y. Li, Y. Liu et J. Zhu. Quantile regression in reproducing kernel hilbert spaces. *Journal of the Americal Statistical Association*, 102 :255–268, 2007.

H.-G. Müller et F. Yao. Functional additive models. *Journal of the Americal Statistical Association*, 103 :426–437, 2008.

C. Preda. Regression models for functional data by reproducing kernel hilbert spaces methods. *Journal of Statistical Planning and Inference*, 137 :829–840, 2007.

F. Rossi et N. Villa-Vialaneix. Support vector machine for functional data classification. *Neurocomputing*, 69 :730–742, 2006.

D. Ruppert et J. Caroll. *Transformation and weighting in regression.* Chapman and Hall, New York, 1988.

B. Schölkopf et A.J. Smola. *Learning with Kernels.* MIT Press, 2002.

I. Steinwart. On the influence of kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2 :67–93, 2001.

I. Steinwart et A. Christmann. *Support Vector Machines.* Springer, New York, 2008.

V. Vapnik. *Statistical Learning Theory.* John Wiley and Sons, 1998.

D-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3) : 739–767, 2002.

Chapitre 7

Perspectives générales

Contents

Ce chapitre tient lieu de conclusion et présente les grands axes de mes recherches futures, dans lesquelles les données fonctionnelles seront toujours présentes.

7.1 Modèles à variables latentes pour données fonctionnelles

Les modèles à variables latentes sont des modèles que je souhaite explorer. De nombreux développements ont été réalisés dans le cadre multivarié (voir Bartholomew et al. (2011), Droesbeke et al. (2013)). Ces modèles postulent l'existence de variables inobservables directement, mais en faisant l'hypothèse que les variables observées sont expliquées par ces variables latentes. La terminologie habituelle des différents modèles à variables latentes dépend de la nature des variables observées et latentes (voir Tableau 7.1).

L'estimation des paramètres dans un modèle à variables latentes peut être basée sur la méthode du maximum de vraisemblance. Cependant, en présence de données

		Variables latentes	
		Qualitatives	Quantitatives
Variables observées	Qualitatives	Modèles de classes latentes	Modèles de traits latents
	Quantitatives	Modèles de profils latents	Modèles d'équations structurelles

Table 7.1 Terminologie des modèles à variables latentes.

fonctionnelles, un problème majeur apparaît : il n'est pas possible de définir la notion de densité de probabilité dans le contexte de données fonctionnelles. Cependant, un travail important de Delaigle et Hall (2010) a permis de définir une notion de pseudo-densité pour des données fonctionnelles. En considérant la décomposition de Karhunen-Loève d'une variable aléatoire fonctionnelle X, sous la forme

$$X = \sum_{j=1}^{+\infty} \sqrt{\lambda_j} \xi_j v_j, \tag{7.1}$$

où $(\lambda_j)_{j\geq 1}$ est la suite décroissante des valeurs propres de l'opérateur de covariance de X et $(v_{j]})_{j\geq 1}$ est la suite associée des fonctions propres, il est possible de définir une suite de densités $(f_j)_{j\geq 1}$ des scores $(\xi_j)_{j\geq 1}$. Pour $h > 0$ et pour $x \in L^2$, nous considérons la probabilité de petites boules $p(x|h) = \mathbb{P}(\|X - x\| \leq h)$. Un résultat donné dans Delaigle et Hall (2010) relie alors ces probabilités de petites boules à la suite de densités $(f_j)_{j\geq 1}$ sous la forme

$$\log p(x|h) = C_1(k, \lambda) + \sum_{j=1}^{k} \log f_j(x_j) + o(k), \tag{7.2}$$

où C_1 est une constante qui ne dépend que de l'ordre k et de la suite de valeurs propres $(\lambda_j)_{j\geq 1}$ et $(x_j)_{j\geq 1}$ est la suite des scores de la variable x dans la base de fonctions propres $(v_{j]})_{j\geq 1}$. Ce résultat permet de définir la log-densité

$$\ell(x|k) = \frac{1}{k} \sum_{j=1}^{k} \log f_j(x_j), \tag{7.3}$$

qui rend compte de l'effet de variation du logarithme de la probabilité de petite boule $p(x|h)$ jusqu'à un certain ordre k correspondant au nombre de valeurs propres considérées.

Cette notion de pseudo-densité a déjà été utilisée par Jacques et Preda (2014) pour réaliser de la classification de courbes, ce qui est un cas particulier de modèle à variables latentes. Cette méthodologie pourrait être adaptée à d'autres modèles pour variables latentes, comme par exemple les modèles à équations structurelles. À l'heure actuelle, nous avons trouvé un seul travail récent Luo et al. (2019) traitant ce type de modèles avec des données fonctionnelles.

7.2 Étude de données éoliennes

7.2.1 Présentation

Dans cette section, je présente un axe de recherche qui a une origine très appliquée, l'étude de données éoliennes pour la prévision verticale de la vitesse du vent. L'énergie

éolienne est une source d'énergie renouvelable, en d'autres termes, il s'agit d'une source d'énergie dont le renouvellement est assez rapide pour qu'elle puisse être considérée comme inépuisable à l'échelle humaine. La maîtrise de ce type d'énergie constitue donc un enjeu majeur. Dans le cas de l'énergie éolienne, la clé de cette maîtrise est de pouvoir prévoir la vitesse du vent à une certaine altitude sur un site donné, pour pouvoir anticiper l'énergie qui sera alors produite par l'éolienne.

Le travail que je présente ici est le fruit d'une collaboration entamée avec Engie Green, filiale d'Engie qui est une entreprise leader des énergies renouvelables en France. Cette collaboration a commencé par l'encadrement du stage de Master 2 de Nizar Soilihi, soutenu le 30 août 2019 (voir Soilihi (2019)).

Par exemple, sur un site donné, des appareils (anémomètres, girouettes, ...) placés sur le mât effectuent des mesures toutes les 10 minutes :

- vitesse du vent à 20, 40, 60, 80, 97, 99 mètres,
- direction du vent à 20, 40, 60, 80, 97, 99 mètres,
- température au sol et à 99 mètres,
- pression atmosphérique au sol et à 99 mètres,
- pourcentage d'humidité au sol et à 99 mètres.

En parallèle de ces mesures, un Lidar (acronyme en anglais de LIght Detection And Ranging) est positionné sur le site pendant une certaine durée (plusieurs mois). Le Lidar est un appareil de mesure à distance qui, grâce à un faisceau lumineux et l'effet Doppler, permet d'effectuer des mesures à des hauteurs plus élevées, par exemple (toutes les 10 minutes) :

- vitesse du vent à 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150 mètres,
- direction du vent à 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150 mètres.

La Figure 1.1 montrée en introduction est un exemple de 30 courbes journalières de la vitesse du vent mesurée par le mât (en mètres par seconde) à 99 mètres d'altitude en un certain lieu. Les mesures étant faites toutes les 10 minutes, nous disposons de 144 points de mesure par courbe. Une autre façon de considérer ces données est de voir la vitesse du vent comme une fonction de l'altitude. La Figure 7.1 est un exemple de 30 courbes de la vitesse du vent mesurée par le mât en fonction de la hauteur (ce sont donc 30 temps d'observation).

L'objectif est de prédire la vitesse à une certaine hauteur élevée (par exemple 120 mètres) en connaissant la courbe de vitesse du vent de 20 à 99 mètres. Nous avons considéré dans un premier temps un modèle linéaire fonctionnel s'écrivant

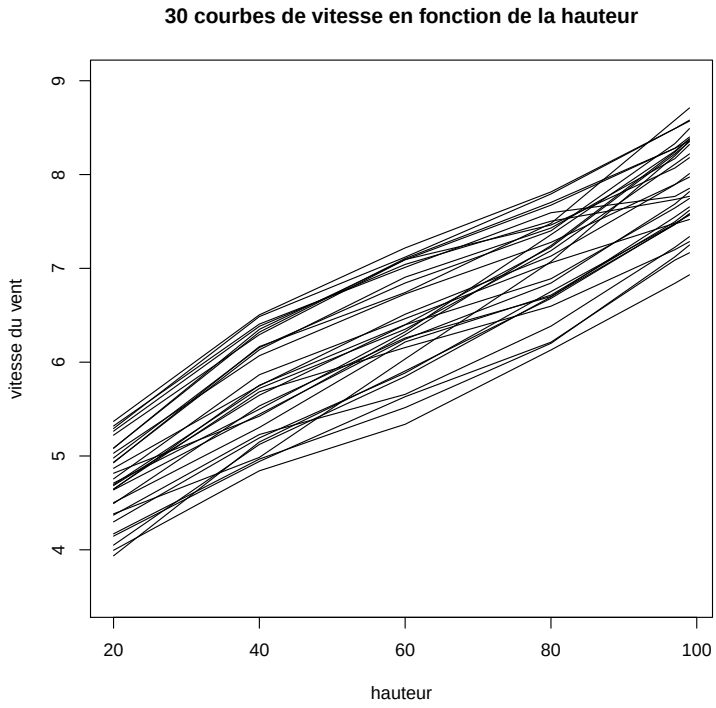

Figure 7.1 Représentation de 30 temps d'observation de la vitesse du vent (en mètres par seconde) en fonction de la hauteur.

$$VV^{120} = \langle VV, \theta \rangle + \varepsilon, \tag{7.4}$$

où VV^{120} désigne la vitesse du vent à 120 mètres d'altitude, VV désigne la courbe de vitesse du vent de 20 à 99 mètres, θ est une fonction inconnue à estimer et ε est le bruit du modèle. Pour estimer θ, nous disposons d'un échantillon d'apprentissage $\left(VV_i, VV_i^{120}\right)_{i=1,\dots,n}$ où n est le nombre de temps d'observation où le Lidar est placé sur le site (en pratique, n est de l'ordre de plusieurs dizaines de milliers). Durant le stage de Master 2 de Nizar Soilihi, différents points ont été abordés :

- mise en œuvre d'une méthode d'estimation de type ridge dans un modèle linéaire fonctionnel pour prédire la vitesse du vent à une altitude (par exemple 120 mètres) en utilisant la courbe de vitesse du vent mesurée de 20 à 99 mètres d'altitude,
- implémentation de l'estimateur dans R pour une utilisation future à Engie Green,
- introduction de nouvelles covariables (température, pression, humidité) dans le modèle,
- comparaison de la qualité de prévision avec les méthodes actuellement utilisées à Engie Green,
- travail sur la recherche d'une période propice à la mise en place du Lidar sur le site.

Dans le prolongement de ce travail, plusieurs points sont amenés à être considérés, que nous allons décrire dans la suite.

7.2.2 Données manquantes

Comme cela a été signalé, l'apparition de données manquantes est incontournable dès que l'on travaille sur des données réelles. Les données éoliennes n'échappent pas à cette règle. Jusqu'à présent, ces données manquantes ont été supprimées. Cette solution simple provoque une perte d'information qu'il pourrait être intéressant de reconstituer le mieux possible. Des méthodes de reconstruction ont déjà été évoquées au chapitre 2. Nous envisageons ici d'étudier une autre méthodologie afin de reconstituer des portions de courbes manquantes, inspirée de la méthode d'imputation "hot deck" (voir Ford (1983)). Pour une certaine courbe présentant une partie manquante, nous sélectionnons un certain nombre de courbes complètes, proches (au sens d'une certaine norme) de celle à reconstituer. Parmi, ces courbes sélectionnées, une courbe peut être tirée aléatoirement pour imputer la partie manquante. Cette procédure peut être répétée un certain nombre de fois. Une autre approche pourrait aussi consister à calculer une moyenne pondérée des courbes sélectionnées. Le fait que les courbes appartiennent à un espace

de dimension infini le rend compliqué à explorer, cela peut être un obstacle à la mise en place de cette méthode "hot deck" sur des données fonctionnelles. L'intérêt de l'étude serait aussi de voir si l'obstacle peut être contourné. Enfin, il pourrait être intéressant de prendre en compte d'autres éléments des courbes (dérivée, ...) qui permettraient d'améliorer la reconstruction.

7.2.3 Données circulaires

Parmi les données mesurées, la direction du vent est une variable importante, qu'il serait intéressant d'introduire dans le modèle. Cette variable présente la particularité d'être circulaire. Plusieurs études ont été réalisées sur ce type de variables dans le cadre réel ou multivarié (voir par exemple Jammalamadala et SenGupta (2001)), nous n'avons rien trouvé à ce jour dans la littérature pour des données fonctionnelles.

Bibliographie

D.J. Bartholomew, M. Knott et I. Moustaki. *Latent variable models and factor analysis : a unified approach, 3rd Edition.* Wiley Series in Probability and Statistics, John Wiley and Sons, Chichester, 2011.

A. Delaigle et P. Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38 :1171–1193, 2010.

J.-J. Droesbeke, G. Saporta et C. Thomas-Agnan. *Modèles à variables latentes et modèles de mélange.* Technip, 2013.

B. Ford. An overview of hot deck procedures. In I. Olkin W. Madow, H. Nisselson, editor, *Incomplete data in sample surveys : Theory and bibliography.* 1983.

J. Jacques et C. Preda. Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis*, 71 :92–106, 2014.

S.R. Jammalamadala et A. SenGupta. *Topics in circular statistics.* Series on multivariate analysis, World Scienti
c, Singapore, 2001.

S. Luo, R. Song, M. Styner, J.H. Gilmore et H. Zhu. Fsem : Functional structural equation models for twin functional data. *Journal of the Americal Statistical Association*, 114 :344–357, 2019.

N. Soilihi. Analyse de données éoliennes et prévision de la vitesse du vent. Master's thesis, 2019.

RÉSUMÉ

Mes recherches portent sur la statistique fonctionnelle. Cette branche de la statistique étudie des modèles lorsque les données sont assimilables à des courbes. En effet, avec les moyens technologiques actuels, certaines données sont relevées sur des grilles de mesure très fines. De plus, ces données proviennent fréquemment de mesures de phénomènes de nature continue (par exemple des relevés de températures, des courbes de croissance, ...), et il est naturel de les traiter en tant que fonctions (du temps, de l'espace) plutôt qu'en tant que vecteurs de points de mesures. On considère alors que l'on observe des réalisations de variables aléatoires à valeurs dans un certain espace de fonctions.

Mes recherches se sont essentiellement focalisées sur la régression mettant en jeu des variables fonctionnelles. Le point de départ est le modèle le plus simple, le modèle de régression linéaire fonctionnelle, dans lequel la variable d'intérêt est à valeurs réelles, et la variable explicative est à valeurs dans un espace de fonctions. Je me suis alors intéressé à différentes extensions de ce modèle.

La première extension introduit des données manquantes dans les observations. Dans un premier temps, je me suis intéressé à une méthode d'imputation de données manquantes sur la variable réponse, afin de reconstituer un jeu de données complet. Ce jeu de données complété permettra une prévision ultérieure lorsque qu'apparaît une nouvelle observation de la variable explicative. Des prolongements de ce travail sont envisagés, notamment le fait de considérer des données manquantes à la fois sur la variable explicative et la variable d'intérêt, ainsi qu'une variable d'intérêt elle aussi à valeurs dans un espace de fonctions.

La deuxième extension considère une variable réponse à valeurs dans un espace de fonctions. Ce travail introduit une méthode d'estimation du paramètre fonctionnel du modèle et établit notamment des résultats asymptotiques sur l'erreur de prévision sur la réponse. En parallèle, je me suis également intéressé à un sous-modèle du modèle précédent dans le cas (notamment lorsque les variables sont fonctions du temps) où la valeur de la variable réponse en un certain instant est expliquée par le passé de la variable explicative avant cet instant.

La troisième extension consiste à se placer dans un cadre non-paramétrique. Deux volets sont abordés dans ce contexte. Le premier, dans le cas de l'estimation de la moyenne conditionnelle, a permis de développer une famille d'estimateurs récursifs à noyau, présentant l'avantage d'être calculés de façon itérative. Le second développe le cas de la régression sur quantiles via la méthode Support Vector Machine (SVM) qui est une méthode d'apprentissage performante basée sur la minimisation d'un risque pénalisé dans un espace de Hilbert à noyau reproduisant.