Modélisation statistique pour données fonctionnelles : approches non-asymptotiques et méthodes adaptatives

Angelina Roche

Collège Doctoral du Languedoc-Roussillon

THÈSE

Pour obtenir le grade de

Docteur

Délivré par l'**Université Montpellier II**

Préparée au sein de l'école doctorale **I2S**
Et de l'unité de recherche **I3M**

Spécialité: **Biostatistique**

Présentée par **Angelina Roche**

Modélisation statistique pour données fonctionnelles : approches non-asymptotiques et méthodes adaptatives.

Soutenue le 7 juillet 2014 devant le jury composé de

Élodie Brunel	Université Montpellier II	Directrice
Fabienne Comte	Université Paris Descartes	Présidente
Christophe Giraud	Université Paris-Sud et École Polytechnique	Examinateur
Jan Johannes	Université Catholique de Louvain	Rapporteur
André Mas	Université Montpellier II	Directeur
Patricia Reynaud-Bouret	Université de Nice Sophia-Antipolis	Rapporteur
Nicolas Verzelen	INRA- SupAgro	Examinateur

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Cette thèse a été préparée à
l'Institut de Mathématiques et de Modélisation de Montpellier (I3M)
Université Montpellier 2
Case courrier 051
34095 Montpellier cedex 05
au sein de l'équipe de Probabilité et Statistiques (EPS).

Remerciements

En premier lieu, je souhaiterais bien sûr remercier mes deux directeurs de thèse, Élodie Brunel et André Mas, pour m'avoir proposé un sujet soulevant des problématiques très intéressantes et pour avoir exaucé mon souhait initial de travailler sur un sujet de recherche mêlant à la fois probabilités et analyse fonctionnelle. Élodie, depuis notre première rencontre, fin 2010, puis durant mon stage de M2 et durant ces trois années de thèse, tu as toujours su te montrer disponible et ta présence a constitué un soutien indispensable. Je te remercie pour cela. Merci à André pour la confiance que tu m'as accordée dès le début de ma thèse et pour avoir été un appui irremplaçable en particulier sur les aspects les plus techniques des résultats présentés dans cette thèse.

Je remercie mes rapporteurs Jan Johannes et Patricia Reynaud, pour la patience dont ils ont fait preuve à la lecture de ce manuscrit, la bienveillance qu'ils ont eue pour les résultats qui y sont présentés et les nombreux conseils donnés pour l'amélioration de ce manuscrit. Jan, merci d'avoir fait le déplacement depuis la Belgique pour assister à ma soutenance et pour ta gentillesse et tes encouragements lors de la période qui l'a précédée. Patricia, merci pour l'intérêt que tu as porté à mes travaux et pour avoir su me mettre en confiance lors des différentes discussions que nous avons eues durant ces dernières semaines. Merci à Fabienne Comte et Nicolas Verzelen, pour avoir suivi ma thèse dès le début, en particulier au travers des comités de suivi de thèse, et pour avoir accepté sans hésitation de faire partie de mon jury. C'est une chance pour moi d'avoir pu interagir avec vous et vos nombreux conseils ont été d'une importance primordiale dans la réussite de cette thèse. Je remercie également Christophe Giraud, dont la richesse et la diversité des travaux scientifiques force bien évidemment mon admiration, de me faire l'honneur de participer à mon jury.

Merci à Gaëlle pour cette collaboration qui s'est révélée très enrichissante et stimulante. Tes connaissances à la fois sur la sélection de modèle et la méthode de Lepski ainsi que tes aptitudes pédagogiques m'ont été d'un grand secours dès le début (n'oublie pas que mes premiers pas dans la recherche se sont basés sur ton mémoire de master !). Merci également pour ton amitié et ton soutien. J'espère que nous aurons l'occasion d'explorer ensemble de nombreuses autres problématiques.

Merci aux membres de l'I3M, en particulier aux membres de l'équipe EPS, à Christian Lavergne, pour m'avoir accueillie si chaleureusement. Un grand merci à Sophie Cazanave-Pin, Bernadette Lacan, Myriam Debus, Éric Hugounenq et Nathalie Quintin pour leur disponibilité, leur sympathie et leur aide précieuse. Je remercie également Catherine Trottier, Gwladys Toulemonde, Christophe Crambes, Bénédicte Fontez, Jean-Michel Marin, Rémi Carles, Jean-Noël Bacro pour leur gentillesse. Merci à Pierre Pudlo pour la confiance que tu m'a accordée pour les TD, les TP et LE cours magistral d'analyse de données et merci pour ton soutien durant la dernière période. Merci à Mathieu Ribatet et Lionel Cucala pour m'avoir guidée dans mes premiers pas d'enseignante. Merci à Gemma Bellal qui a toujours un petit mot gentil à mon égard. Merci à Vivianne Durand-Guerrier pour tes encouragements. Merci à Simon Mendez et Vanessa Lleras pour leur sympathie et leur humour. Un très grand merci à Baptiste pour ta disponibilité, ton aide logistique et ta patience lors de mes hésitations pour le choix de la salle de soutenance.

Je souhaite remercier également tout les doctorants, ex-doctorants et ATER du labo, en particulier mes co-bureaux : Jojo (qui m'a aidé à m'orienter dans les différentes démarches administratives de début de thèse), Christophe (cela aura été un plaisir de partager ce bureau avec toi pendant ces trois ans), Étienne (avec qui j'ai eu la chance de co-organiser le séminaire des doctorants et désolée de ne pas être arrivée aux 100 paniers !) et Bastien. Mais aussi tout les autres avec qui j'ai passé de très agréables moments : Afaf, Arnaud, Benjamin, Claudia, Coralie, Christian, Damien B., Damien J., Daria, Elsa, Guillaume, Jean, Julianna, Julien le jeune, Julien le stateux, Junior, Mathieu C., Mathieu L., Mickaël, Myriam, Thomas, Tutu, Victor, Vincent, Yousri (the King),... ainsi que ceux du LMGC en particulier Adrien, François, Li, Nawfal, Paul, Rémi,... Un grand merci à Dalia, Pauline, Claire, David, Léa, Isa et Laeticia pour tous les bons moments que nous avons passés ensemble.

An important part of this document has been written in the University of Copenhagen. I want to express my sincere thanks to the members of the Statistics and Probability theory group for hosting me during more than two months. In particular, I would like to thanks Susanne Ditlevsen, which has been very kind to me, Helle Sørensen, Anders Tolver and Bo Markussen. I hope that we will find other occasions to work together. I also want to thanks the people I have met during my stay : Giacomo, Jost, Lars Lau, Mathias, Nadim, Nourah, Robin, Sima, Tomasz,...

Un grand merci à ma famille sans qui, bien sûr, je ne serai pas là. Un grand merci en particulier à mes grand-parents, Florine et Yves, qui ont mis de l'argent de côté dès ma naissance pour que je puisse faire des études et qui n'ont cessé de me soutenir. Un grand merci à ma mère, pour son amour inconditionnel, à ma cousine Pauline qui a toujours été une sœur pour moi, à mes oncles Roland et Christian qui m'ont toujours encouragée. Merci à Marie, qui m'a tendu la main à un moment où j'en avais grand besoin. Merci à mon père, à ma grand-mère Christiane et à mes tantes Nathalie et Laura pour leur présence aujourd'hui et leur soutien. Je remercie également Alain et Christiane pour avoir renoncé à la Cocarde d'Or pour venir ici aujourd'hui. Un grand merci à Lucie et Émilie pour leurs encouragements de l'autre côté de l'Atlantique.

Enfin, mes plus tendres et amoureux remerciements vont à Vincent, pour tout le chemin que nous avons parcouru ensemble depuis bientôt dix ans. Pour ta présence au quotidien et ton soutien infaillible dans les meilleurs moments, comme dans les pires, malgré l'éloignement et les soucis. Après un tour de France et un passage au Danemark, j'espère que nous trouverons enfin bientôt un endroit où nous poser durablement.

Mes derniers remerciements vont également à ceux qui ont permis de mettre en place et de conserver un système d'enseignement qui offre encore à tous, quelle que soit son origine sociale, un accès à un enseignement de grande qualité. Sans cela, sans les bourses, sans le statut de normalien qui m'a permis d'être très tôt autonome financièrement, je ne serai pas arrivée au bout de ces neuf années d'étude.

Organisation de la thèse

Une partie importante des travaux présentés dans ce mémoire consiste à proposer de nouvelles procédures d'estimation en statistique pour données fonctionnelles permettant d'obtenir des estimateurs adaptatifs dont les propriétés sont étudiées d'un point de vue non-asymptotique. Une implémentation pratique des estimateurs est également proposée. Un autre aspect de ces travaux de thèse a été de s'intéresser à des problèmes d'optimisation en adaptant au cadre fonctionnel la méthodologie des surfaces de réponse.

Le chapitre 1 est un chapitre introductif présentant le contexte scientifique de la thèse et ses principales contributions.

Nous présentons dans le chapitre 2 une procédure d'estimation adaptative dans le cadre du modèle linéaire fonctionnel. Les estimateurs proposés sont des estimateurs par projection dont la dimension est sélectionnée par un critère des moindres carrés pénalisé.

Le chapitre 3 est consacré à un travail réalisé en collaboration avec Gaëlle Chagny (LMRS) sur l'estimation de la fonction de répartition d'une variable Y conditionnellement à une variable fonctionnelle X. Nous nous intéressons ici à des estimateurs à noyau et nous proposons un critère de sélection de la fenêtre dans l'esprit de la méthode dite de *Goldenshluger-Lepski*.

Le chapitre 4 est consacré à la méthode des surfaces de réponse, communément utilisée dans l'industrie, pour développer des plans d'expériences *guidés* dans des contextes multivariés. Nous motivons et proposons une manière d'étendre cette méthode à un cadre fonctionnel.

Nous présentons dans l'annexe A la théorie de la perturbation qui est un outil indispensable pour prouver les résultats principaux du chapitre 2.

Enfin, l'annexe B donne les énoncés des inégalités de concentration utilisées tout au long de la thèse.

Les chapitres 1, 2, 3 et 4 peuvent être lus indépendamment les uns des autres.

Table des matières

Notations

$\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$	ensemble des entiers naturels, entiers relatifs, nombres rationnels, nombres réels et nombres complexes
$\mathbb{N}^*$	ensemble des entiers naturels non nuls : $\mathbb{N}^* = \mathbb{N}\backslash\{0\}$
$(\Omega, \mathcal{A}, \mathbb{P})$	espace de probabilité : Ω ensemble non-vide, $\mathcal{A}$ σ-algèbre formée de sous-ensembles de Ω, $\mathbb{P}$ mesure de probabilité sur $\mathcal{A}$
$\mathbb{E}[Z]$, $\mathrm{Var}(Z)$	espérance, variance d'une variable aléatoire réelle Z
$(\mathbb{H}, \langle\cdot,\cdot\rangle, \|\cdot\|)$	espace de Hilbert séparable réel muni d'un produit scalaire $\langle\cdot,\cdot\rangle$ et d'une norme associée $\|x\| = \sqrt{\langle x, x\rangle}$
$\mathbb{H}^*$	dual topologique de $\mathbb{H}$
$\dim(E)$	dimension de l'espace vectoriel E
X	variable aléatoire à valeurs dans $\mathbb{H}$
$\mathbb{E}[X]$	espérance de X (voir p. 4)
P_X	loi de X
$X_1, \ldots, X_n$	suite de variables aléatoires i.i.d.(indépendantes et identiquement distribuées) suivant la même loi que X
$\mathbf{X}$	$(X_1, \ldots, X_n)$
$\mathbb{E}_{\mathbf{X}}$, $\mathbb{P}_{\mathbf{X}}$	espérance conditionnelle, probabilité conditionnelle par rapport à $\mathbf{X}$
Γ	opérateur de covariance associé à X (Éq. (1.3), p.4)
Γ_n	opérateur de covariance empirique calculé à partir de $X_1, \ldots, X_n$ (Éq. (1.5), p.5)
$\Gamma^{1/2}$	racine carrée de l'opérateur Γ : $\Gamma^{1/2}\Gamma^{1/2} = \Gamma$
$(\psi_j)_{j\geq 1}$, $(\lambda_j)_{j\geq 1}$	fonctions propres et valeurs propres de Γ
$(\widehat{\psi}_j)_{j\geq 1}$, $(\widehat{\lambda}_j)_{j\geq 1}$	fonctions propres et valeurs propres de Γ_n
$\mathrm{Ker}(\Gamma)$	noyau de l'opérateur Γ
$\mathrm{Im}(\Gamma)$	image de l'opérateur Γ
$\mathrm{rg}(\Gamma)$	rang de l'opérateur Γ : $\mathrm{rg}(\Gamma) = \dim(\mathrm{Im}(\Gamma))$
$\mathrm{Vect}\{x_1, \ldots, x_D\}$	espace vectoriel engendré par $\{x_1, \ldots, x_D\}$
$\mathrm{Ker}(\Gamma)$	noyau de l'opérateur Γ
$\mathcal{N}(\mu, \sigma^2)$	loi normale de moyenne $\mu \in \mathbb{R}$ et de variance $\sigma^2 > 0$
$\mathcal{U}([a, b])$	loi uniforme sur l'intervalle $[a, b]$
$a \wedge b$, $a \vee b$	minimum et maximum de deux nombres réels a et b
$\mathbf{1}_A$	fonction indicatrice d'un ensemble A : $\mathbf{1}_A(x) = 1$ si $x \in A$ et $\mathbf{1}_A(x) = 0$ si $x \notin A$
$\mathrm{Re}(z)$, $\mathrm{Im}(z)$	parties réelle et imaginaire d'un nombre complexe z

Notation

$\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$	set of natural numbers, integers, rational numbers, real numbers and complex numbers
$\mathbb{N}^*$	set of positive natural numbers: $\mathbb{N}^* = \mathbb{N}\backslash\{0\}$
$(\Omega, \mathcal{A}, \mathbb{P})$	probability space: Ω non empty set, $\mathcal{A}$ σ-algebra of subsets of Ω, $\mathbb{P}$ probability measure on $\mathcal{A}$
$\mathbb{E}[Z]$, $\mathrm{Var}(Z)$	expectation, variance of a real random variable Z
$(\mathbb{H}, \langle\cdot,\cdot\rangle, \|\cdot\|)$	real Hilbert space equipped with a scalar product $\langle\cdot,\cdot\rangle$ and its associated norm $\|x\| = \sqrt{\langle x, x\rangle}$
$\mathbb{H}^*$	topological dual of $\mathbb{H}$
$\dim(E)$	dimension of the vector space E
X	random variable taking values in $\mathbb{H}$
$\mathbb{E}[X]$	expectation of X (see p. 4)
P_X	distribution of X
$X_1, \ldots, X_n$	sequence of i.i.d. (independent and identically distributed) random variables following the same distribution as X
$\mathbf{X}$	$(X_1, \ldots, X_n)$
$\mathbb{E}_{\mathbf{X}}$, $\mathbb{P}_{\mathbf{X}}$	conditional expectation, probability with respect to $\mathbf{X}$
Γ	covariance operator of X (Éq. (1.3), p.4)
Γ_n	empirical covariance operator calculated from $X_1, \ldots, X_n$ (Éq. (1.5), p.5)
$\Gamma^{1/2}$	square root of the operator Γ : $\Gamma^{1/2}\Gamma^{1/2} = \Gamma$
$(\psi_j)_{j\geq 1}$, $(\lambda_j)_{j\geq 1}$	eigenfunctions and eigenvalues of Γ
$(\widehat{\psi}_j)_{j\geq 1}$, $(\widehat{\lambda}_j)_{j\geq 1}$	eigenfunctions and eigenvalues of Γ_n
$\mathrm{Ker}(\Gamma)$	kernel of the operator Γ
$\mathrm{Im}(\Gamma)$	image of the operator Γ
$\mathrm{rg}(\Gamma)$	rank of the operator Γ: $\mathrm{rg}(\Gamma) = \dim(\mathrm{Im}(\Gamma))$
$\mathrm{span}\{x_1, \ldots, x_D\}$	vector space spanned by $\{x_1, \ldots, x_D\}$
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution of mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$
$\mathcal{U}([a, b])$	uniform distribution on the interval $[a, b]$
$a \wedge b$, $a \vee b$	minimum and maximum of two real numbers a and b
$\mathbf{1}_A$	indicator function of a set A : $\mathbf{1}_A(x) = 1$ if $x \in A$ and $\mathbf{1}_A(x) = 0$ if $x \notin A$
$\mathrm{Re}(z)$, $\mathrm{Im}(z)$	real and imaginary parts of a complex number

Chapitre 1

Contexte scientifique et contributions

Nous présentons dans ce chapitre le domaine statistique dans lequel s'insère cette thèse ainsi que les principaux résultats présentés dans la suite du manuscrit. Le contexte scientifique est celui des données fonctionnelles, plus précisément nous nous intéresserons à des problèmes d'estimation dans des modèles liant une variable d'intérêt réelle Y et une covariable fonctionnelle X. Dans l'objectif de proposer des méthodes d'estimation adaptatives dans un cadre non-asymptotique, nous nous sommes appuyés sur des outils de sélection de modèle et de sélection de fenêtre que nous présentons dans la section 1.2. Le premier travail de cette thèse porte sur une procédure d'estimation dans le cadre du modèle linéaire fonctionnel défini en section 1.3. Nous présentons ensuite les problématiques liées à l'estimation des fonctionnelles de régression en section 1.4 et la méthode des surfaces de réponses en section 1.5. Les contributions de cette thèse sont décrites en sections 1.3.4, 1.4.2 et 1.5.3.

Sommaire

1.1 Données fonctionnelles

L'analyse des données fonctionnelles (Ramsay et Silverman, 2005 ; Ferraty et Vieu, 2006 ; Ferraty et Romain, 2011) a connu un intérêt croissant durant les vingt dernières années, intérêt motivé par de nombreuses applications. En effet, les progrès techniques récents permettent d'enregistrer des données sur des grilles de plus en plus fines. Typiquement, les données sont récoltées sous la forme suivante $(X_i(t_{i,1}), \dots, X_i(t_{i,p}))_{1 \leq i \leq n}$ où $(t_{i,1}, \dots, t_{i,p})$ est une suite ordonnée (par exemple une discrétisation temporelle) et n la taille de l'échantillon.

L'apport de l'analyse des données fonctionnelles consiste à traiter ce type de données comme une séquence de réalisations $(X_i)_{1 \leq i \leq n}$ d'une variable aléatoire à valeurs dans un espace de fonctions et observée à certains instants. L'intérêt de cette approche est double : d'une part, il est assez fréquent que p soit très grand ($p >> n$) ce qui rend ce type de données très difficiles (voire impossibles) à traiter avec les méthodes classiques de la statistique multivariée. D'autre part, il arrive souvent en pratique que les données ne soient pas récoltées sur la même grille, par exemple lorsque certaines observations sont manquantes. L'analyse des données fonctionnelles offre un cadre théorique et numérique utile pour résoudre ce genre de problème.

Les applications pratiques de cette approche sont de plus en plus nombreuses, parmi les plus récentes, nous pouvons citer : l'étude d'électroencéphalogrammes (Di et al., 2009), l'analyse des mouvements en biomécanique (voir par exemple Sørensen et al. 2012), l'étude de l'évolution des taux d'intérêts au cours du temps (Laurini, 2014), l'étude des audiences télévisuelles (Cardot, Cénac et Zitt, 2012), des profils de croissance (Sauder et al., 2013),... L'analyse des données fonctionnelles ne se limite pas à étudier des quantités évoluant au cours du temps : par exemple des données spatiales (Rakêt et Markussen, 2014) peuvent également être considérées.

Nous nous intéresserons ici uniquement à la modélisation et à l'étude du lien entre une variable aléatoire fonctionnelle X et une variable aléatoire réelle Y. La variable X sera supposée à valeurs dans un espace de Hilbert séparable $(\mathbb{H}, \langle \cdot, \cdot \rangle, \| \cdot \|)$, c'est-à-dire un espace vectoriel muni d'un produit scalaire $\langle \cdot, \cdot \rangle$, complet pour la norme associée $\|x\| = \sqrt{\langle x, x \rangle}$ et admettant un sous-ensemble dénombrable dense (nous renvoyons aux chapitres III.6 et V.1 de Brezis (2005) pour plus de précisions sur ces notions). Typiquement $\mathbb{H} = \mathbb{L}^2(I)$ pour I un intervalle de $\mathbb{R}$ ou un espace de Sobolev. Nous passons donc du cadre fini-dimensionnel de la statistique classique à un cadre infini-dimensionnel. La plupart des outils et notions classiques de statistique multivariée se généralisent au cadre fonctionnel, sous réserve de prendre certaines précautions.

1.1.1 Notions d'espérance et de covariance

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité. Nous munissons l'espace $\mathbb{H}$ de sa tribu borélienne $\mathcal{B}(\mathbb{H})$. Dans la suite X désignera une variable aléatoire à valeurs dans $\mathbb{H}$, c'est-à-dire une application mesurable $X : (\Omega, \mathcal{A}) \to (\mathbb{H}, \mathcal{B}(\mathbb{H}))$. Nous noterons P_X la loi de X.

Espérance

Si $\mathbb{E}[\|X\|] < +\infty$, nous pouvons définir l'espérance de X comme la quantité suivante :

$$\mathbb{E}[X] = \int_{\Omega} X(\omega) d\mathbb{P}(\omega). \tag{1.1}$$

L'intégrale de Lebesgue n'étant pas définie pour des fonctions à valeurs dans des espaces de dimension infinie, l'intégrale écrite ici est ce qu'on appelle l'*intégrale de Bochner* de X par rapport à la mesure $\mathbb{P}$. Elle est définie de manière analogue à l'intégrale de Lebesgue, c'est-à-dire par approximation de X par des fonctions « simples » (pour plus de détails voir Dunford et Schwartz 1958, Chapitre III.2). Notons ici que nous définissons ainsi un élément de $\mathbb{H}$. L'espérance définie de cette manière a de bonnes propriétés, notamment, pour tout $f \in \mathbb{H}$,

$$\mathbb{E}[\langle X, f\rangle] = \langle \mathbb{E}[X], f\rangle,$$

ce qui implique, en choisissant $f = \mathbb{E}[X]$, que

$$\|\mathbb{E}[X]\| \leq \mathbb{E}[\|X\|],$$

et

$$\|\mathbb{E}[X]\|^2 \leq \mathbb{E}[\|X\|^2].$$

Nous dirons que X est *centrée* lorsque $\mathbb{E}[X] = 0$.

Une autre manière possible de définir l'espérance de X, lorsque cela a du sens, est de prendre l'espérance « point à point » de X. Bien que ces deux notions d'espérance soient de nature complètement différente, elles coïncident souvent. Par exemple, lorsque $\mathbb{H} = \mathbb{L}^2(I)$, avec I un intervalle de $\mathbb{R}$, s'il existe une application mesurable $\mathbb{X} : \Omega \times I \to \mathbb{R}$ telle que, pour tout $\omega \in \Omega$, $X(\omega)$ est la classe dans $\mathbb{L}^2(I)$ de l'application $t \mapsto \mathbb{X}(\omega, t)$, nous pouvons définir $m(t) = \int_{\Omega} \mathbb{X}(\omega, t) d\mathbb{P}(\omega)$ l'espérance « point à point » de X, qui est bien de carré intégrable dès que $\mathbb{E}[\|X\|] < \infty$ et vérifie

$$\mathbb{E}[X] = m \text{ p.p.} \tag{1.2}$$

Covariance

La notion de covariance se généralise également. La spécificité du cadre fonctionnel par rapport au cadre multivarié est que la covariance de X prend la forme d'un opérateur et non d'une matrice.

Nous supposerons dans toute la suite que $\mathbb{E}[\|X\|^2] < +\infty$. Nous définissons l'*opérateur de covariance* de X de la façon suivante

$$\Gamma : f \in \mathbb{H} \mapsto \mathbb{E}[\langle X, f\rangle X]. \tag{1.3}$$

De manière analogue à la matrice de covariance dans le cadre vectoriel qui est symétrique et positive, l'opérateur Γ est auto-adjoint et positif, c'est-à-dire que pour tous $f, g \in \mathbb{H}$,

$$\langle \Gamma f, g\rangle = \langle f, \Gamma g\rangle \text{ et } \langle \Gamma f, f\rangle \geq 0.$$

Il est de plus compact (Bosq, 2000, pp. 36–37) ce qui le rend diagonalisable (Brezis, 2005, Théorème VI.11). Nous noterons $(\psi_j)_{j\geq 1}$ ses fonctions propres (qui forment une base de Hilbert de $\mathbb{H}$) et $(\lambda_j)_{j\geq 1}$ les valeurs propres associées, rangées dans l'ordre décroissant. Notons que, comme Γ est positif, toutes ses valeurs propres sont positives également et, de plus

$$\mathbb{E}[\|X\|^2] = \sum_{j\geq 1} \mathbb{E}[\langle X, \psi_j\rangle^2] = \sum_{j\geq 1} \lambda_j.$$

Comme $\mathbb{E}[\|X\|^2] < +\infty$, nous avons également $\sum_{j\geq 1}\lambda_j < +\infty$ et l'opérateur Γ est donc un opérateur *à trace* (Simon, 2005, Chapitre 3) ou *nucléaire* (Bosq, 2000, p.35).

La connaissance de cet opérateur donne un certain nombre d'informations sur la variable X, en particulier

$$X = \sum_{j\geq 1}\langle X, \psi_j\rangle\psi_j = \sum_{j\geq 1}\sqrt{\lambda_j}\xi_j\psi_j, \tag{1.4}$$

avec $\xi_j := \langle X, \psi_j\rangle/\sqrt{\lambda_j}$ lorsque $\lambda_j > 0$, la convergence de la série ayant lieu en norme $\|\cdot\|$. Lorsque X est centrée, la suite $(\xi_j)_{j\geq 1}$ définit une suite de variables aléatoires centrées, réduites et non-corrélées. L'écriture de X sous la forme (1.4) s'appelle la *décomposition de Karhunen-Loève* de X.

À partir d'un échantillon $X_1, \ldots, X_n$ de même loi que X, nous pouvons définir une version empirique de l'opérateur de covariance

$$\Gamma_n : f \in \mathbb{H} \mapsto \frac{1}{n}\sum_{i=1}^{n}\langle f, X_i\rangle X_i. \tag{1.5}$$

Cet opérateur Γ_n est de rang fini donc également compact, et auto-adjoint. Il est ainsi diagonalisable, nous noterons $(\widehat{\psi}_j)_{j\geq 1}$ ses fonctions propres et $(\widehat{\lambda}_j)_{j\geq 1}$ ses valeurs propres. La suite de valeurs propres $(\widehat{\lambda}_j)_{j\geq 1}$ est nulle à partir d'un certain rang $J_0 := \mathrm{rg}(\Gamma_n)+1 \leq n+1$.

D'autres propriétés des opérateurs Γ et Γ_n seront données dans la Section A.1.1.

Analyse en Composantes Principales (ACP)

L'objectif de l'analyse en composantes principales est de détecter les sous-espaces de dimension finie de $\mathbb{H}$ qui « capturent » le plus d'information sur la variable X.

Le premier travail portant sur l'ACP fonctionnelle est celui de Dauxois, Pousse et Romain (1982), de nombreux travaux ont suivi (voir Hall 2011 et les références citées). De même qu'en statistique multivariée, le premier intérêt de l'ACP est de fournir un outil pour visualiser la répartition des données (Ramsay et Silverman, 2005, Section 8.3) mais l'ACP est également très utilisée pour réduire la dimension des données par exemple en régression linéaire fonctionnelle (nous y reviendrons par la suite dans la section 1.3 et le chapitre 2).

De manière analogue à l'ACP multivariée, nous définissons une base orthonormée $(\psi_j)_{j\geq 1}$ de $\mathbb{H}$ par récurrence. L'objectif étant de trouver un espace $S_k := \mathrm{Vect}\{\psi_1, \ldots, \psi_k\}$ de dimension k minimisant en moyenne la distance $\mathbb{L}^2$ entre X et sa projection sur S_k. Supposons que nous ayons déterminé $\psi_1, \ldots, \psi_k$, alors, notons $\Pi_k X := \sum_{j=1}^{k}\langle X, \psi_j\rangle\psi_j$ la projection orthogonale de X sur $\mathrm{Vect}\{\psi_1, \ldots, \psi_k\}$,

$$\psi_{k+1} \in \arg\min_{f\in\mathbb{H}} \mathbb{E}\left[\|X - \Pi_k X - \langle X, f\rangle f\|^2\right],$$

sous la contrainte $\langle \psi_{k+1}, \psi_j \rangle = 0$ pour tout $j \leq k$ et $\|\psi_{k+1}\| = 1$. L'espace $\mathbb{H}$ étant séparable, la famille $(\psi_k)_{k\geq 1}$ ainsi définie est donc au plus dénombrable. Nous pouvons vérifier également que $(\psi_k)_{k\geq 1}$ est formée des fonctions propres de l'opérateur de covariance Γ rangées par ordre décroissant suivant les valeurs propres associées. Remarquons ici qu'il y a une légère ambiguïté dans la définition de ψ_k, en effet, si ψ_k vérifie la définition alors $-\psi_k$ la vérifie également, nous connaissons donc seulement la base de l'ACP à un signe près.

Il est possible, sous certaines conditions, que la base de fonctions propres de Γ soit connue même si l'opérateur de covariance lui-même est inconnu, c'est le cas par exemple des données *circulaires* (Comte et Johannes, 2010) c'est-à-dire lorsque $X \in \mathbb{L}^2([a,b])$, $X(a) = X(b)$ et X stationnaire au second-ordre. Toutefois, dans le cas général, la base $(\psi_j)_{j\geq 1}$ est inconnue du statisticien. Dans ce cas-là, nous en définissons une version empirique en prenant les fonctions propres $\left(\widehat{\psi}_j\right)_{j\geq 1}$ de l'opérateur de covariance empirique. La convergence et la normalité asymptotique de tels estimateurs ont été étudiées par exemple par Dauxois, Pousse et Romain (1982), Hall et Hosseini-Nasab (2006) et Cardot, Mas et Sarda (2007) souvent sous la forme d'une étude de la convergence des projecteurs sur les espaces Vect $\left\{\widehat{\psi}_1, \ldots, \widehat{\psi}_K\right\}$ ou Vect $\left\{\widehat{\psi}_j\right\}$ vers leur équivalent théorique. De récents travaux (Hilgert, Mas et Verzelen, 2013 ; Mas et Ruymgaart, 2014) ont porté également sur l'étude du comportement de tels projecteurs à taille d'échantillon fixée. S'appuyant sur ces travaux, nous avons montré des résultats de contrôles de ces projecteurs, présentés en Annexe A (voir par exemple Lemme 18 p. 165).

Il est à noter que, si les courbes de l'échantillon sont peu régulières, les estimateurs $(\widehat{\psi}_j)_{j\geq 1}$ le seront également, c'est la raison pour laquelle des versions lissées de tels estimateurs ont été définies (Rice et Silverman, 1991 ; Lee, Zhang et Song, 2002 ; Ramsay et Silverman, 2005).

1.1.2 Une classe importante de processus : les processus gaussiens

Nous nous arrêtons ici sur une classe importante de processus que sont les processus gaussiens. Un processus $X : I \to \mathbb{R}$ est dit gaussien si, pour tout $k \in \mathbb{N}^*$, pour tous $t_1, \ldots, t_k \in I$, le vecteur $(X(t_1), \ldots, X(t_k))$ est gaussien. De manière plus générale, un élément X de l'espace de Hilbert $\mathbb{H}$ est dit gaussien si, pour toute forme linéaire $\Lambda \in \mathbb{H}^*$, ΛX est une variable gaussienne. Ces deux définitions peuvent concorder lorsque cela a un sens, par exemple, si X est un processus gaussien sur I, alors la classe de X dans $\mathbb{L}^2(I)$ est gaussienne (voir Ibragimov et Rozanov 1978, sections 1.2 et 1.4).

Lorsque X est un processus gaussien, les coefficients $(\xi_j)_{j\geq 1}$ apparaissant dans la décomposition de Karhunen-Loève (1.4) de X sont indépendants.

Le processus gaussien le plus connu est sans aucun doute le mouvement brownien standard qui vérifie, de plus, les trois propriétés suivantes.

- La fonction $t \to X(t)$ est presque sûrement continue.
- L'espérance de X est nulle et $\mathbb{E}[X(t)X(s)] = \min\{s,t\}$ pour tous $s,t \in I$.
- $X(0) = 0$ p.s.

La décomposition de Karhunen-Loève du mouvement brownien standard restreint à $[0,1]$

Mouvement brownien $(W(t))_{t\in[0,1]}$ | Pont brownien $(B(t))_{t\in[0,1]}$ | Processus plus régulier $X(t) = \xi_1 + \sqrt{2}\xi_2 \sin(-\pi t/2)$, $\xi_1, \xi_2 \sim \mathcal{N}(0,1)$

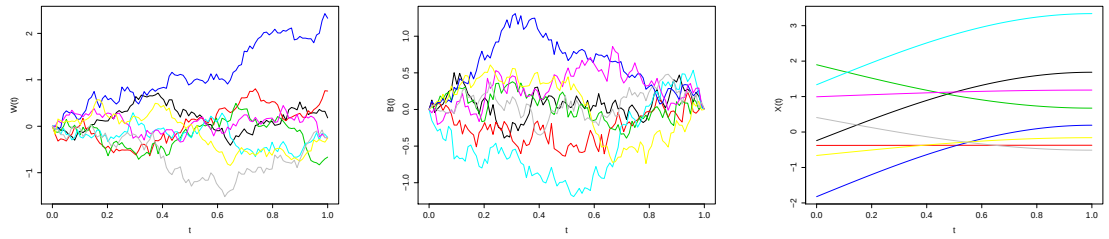


FIGURE 1.1 – Exemples de processus gaussiens.

est connue précisément (voir Ash et Gardner 1975, pp. 41–42) :

$$X(t) = \sqrt{2}\sum_{j\geq 1} \xi_j \frac{\sin(\pi(j-0.5)t)}{\pi(j-0.5)},$$

la convergence ayant lieu presque sûrement dans $\mathbb{L}^2([0,1])$ et uniformément. Nous nous baserons sur cette décomposition pour simuler une large classe de processus X (incluant le mouvement brownien) dans les chapitres 2, 3 et 4.

La décomposition de Karhunen-Loève du pont brownien $B(t) := W(t) - tW(1)$ (MacNeill, 1978 ; Deheuvels, 2007) est donnée par :

$$B(t) = \sum_{j\geq 1} \xi_j \frac{\sqrt{2}\sin(j\pi t)}{j\pi}.$$

1.1.3 La question de la densité

Une des difficultés à traiter des données fonctionnelles réside en l'absence de mesure universellement acceptée comme mesure de référence en dimension infinie. En d'autres termes, il n'y a pas d'équivalent fonctionnel de la mesure de Lebesgue, ce qui ne permet pas de définir la notion de densité aussi aisément qu'en dimension finie.

Par exemple, il ne peut pas exister de mesure non triviale localement finie invariante par translation en dimension infinie. En effet, supposons que μ soit une telle mesure, alors il existe une boule B de rayon $\delta > 0$ telle que $\mu(B) < +\infty$. Comme $\mathbb{H}$ est de dimension infinie, nous pouvons trouver une suite de boules $(B_j)_{j\in\mathbb{N}}$ disjointes deux à deux et de mêmes rayons telles que $\bigcup_{j\in\mathbb{N}^*} B_j \subset B$ [1]. Comme μ est invariante par translation, toutes ces boules ont même mesure, ce qui est contradictoire avec le fait que $\mu(B) < +\infty$ sauf si $\mu = 0$.

Delaigle et Hall (2010) ont proposé de définir une notion de log-densité à partir de la

1. Nous pouvons voir par exemple que, lorsque $(e_j)_{j\geq 1}$ est une base hilbertienne de $\mathbb{H}$ et x_0 est le centre de la boule B, la suite $(B_j)_{j\geq 1}$ des boules ouvertes de centre $x_0 + \delta e_j/(2\sqrt{2})$ et de rayon $\delta/4$ convient. Ce résultat, très proche du théorème de Riesz, est vrai pour tout espace vectoriel normé de dimension infinie et peut également être démontré, de manière plus technique, sans faire appel à la notion de base hilbertienne.

densité de $(\xi_j)_{j=1,\dots,K}$ pour un certain K. Cette définition donne un sens par exemple à la notion de mode et fournit un outil supplémentaire pour visualiser les données.

Il existe des exemples particuliers de familles de processus dont les lois sont absolument continues par rapport à une certaine mesure, qui sert ainsi de mesure de référence. Par exemple, la loi du mouvement brownien translaté W_F, défini par $W_F(t) := W(t) + F(t)$ avec W le mouvement brownien sur $[0,1]$ et F une fonction continue telle que $F(0) = 0$, est absolument continue par rapport à la loi du mouvement brownien si (et seulement si) la fonction F appartient à l'espace de Dirichlet, c'est-à-dire qu'il existe une fonction $f \in \mathbb{L}^2([0,1])$ telle que $F(t) = \int_0^t f(s)ds$. Ce résultat est connu sous le nom de théorème de Cameron-Martin (voir Mörters et Peres 2010, Chapitre 1).
Lorsque le processus X est supposé appartenir à une telle famille, la densité de X par rapport à la mesure de référence est bien définie et il est possible de l'estimer. Dans ce cadre-là, des estimateurs à noyau ont été étudiés par Dabo-Niang (2004) puis généralisés par Dabo-Niang et Yao (2013) pour l'estimation de la densité de processus spatiaux $(X_i, i \in (\mathbb{N}^*)^N)$ avec $N > 1$.

1.2 Méthodes adaptatives en statistique non-paramétrique

Le cœur de cette thèse porte sur des problèmes de sélection d'un estimateur $\tilde{s}$ dans une famille d'estimateurs $\{\widehat{s}_\lambda, \lambda \in \Lambda\}$. Par exemple, le paramètre λ peut être la dimension du sous-espace de projection lorsque l'on considère des estimateurs par projection ou une fenêtre si l'on considère des estimateurs à noyau ou de manière plus large recouvrir un ensemble d'estimateurs de différents types. De manière heuristique, on dit qu'une méthode de sélection est *adaptative* si elle sélectionne un estimateur de manière "optimale", dans un sens à préciser, quels que soient les paramètres (inconnus) du modèle.

Différentes méthodes existent. Sous le nom d'*agrégation* (Nemirovski, 2000 ; Yang, 2003 ; Tsybakov, 2008) est regroupé un ensemble de méthodes consistant à définir un estimateur $\widetilde{s} := \sum_{\lambda\in\Lambda} \theta_\lambda \widehat{s}_\lambda$ comme combinaison linéaire d'éléments de la famille $\{\widehat{s}_\lambda, \lambda \in \Lambda\}$ (Λ est supposée finie) qui s'appelle ici *dictionnaire*. L'élément θ est ensuite souvent sélectionné par minimisation sous contrainte d'un risque pénalisé. Différents types de pénalité ont été introduites : pénalisation ℓ_0 ou BIC (Schwarz, 1978), pénalisation ℓ_1 (LASSO, Tibshirani 1996), pénalisation ℓ_2 (régression ridge, Hastie, Tibshirani et Friedman 2009), elastic net (Zou et Hastie, 2005),... Une autre grande classe de méthodes consiste à sélectionner un estimateur par *validation croisée* (Stone, 1974 ; Allen, 1974 ; Geisser, 1975). Cela consiste à diviser l'échantillon en deux parties : un échantillon d'apprentissage sur lequel les différents estimateurs sont calculés et un échantillon de test qui permet d'estimer le risque de chaque estimateur. La procédure est éventuellement répétée avec un grand nombre d'échantillons d'apprentissage. Cette méthode est très utilisée, notamment car elle s'adapte facilement à des contextes variés. Toutefois, peu de résultats théoriques existent sur cette méthode (voir Arlot et Celisse 2010 pour un état de l'art). Récemment, des travaux permettant d'évaluer, d'un point de vue non-asymptotique, la qualité d'un estimateur sélectionné par validation croisée ont été proposés, dans des contextes d'estimation de densité (Celisse, 2008 ; Arlot et Lerasle, 2012) et de régression (Arlot, 2008).

Lorsque les estimateurs de la famille $\{\widehat{s}_\lambda, \lambda \in \Lambda\}$ sont de même nature, l'objectif peut

être de sélectionner un seul estimateur. La méthode de sélection de modèle par contraste pénalisé et la méthode plus récente de Goldenshluger et Lepski (2011) consistent toutes deux à sélectionner le paramètre $\lambda \in \Lambda$ par un critère imitant la décomposition biais-variance du risque de l'estimateur. Nous développons plus en détail le principe de ces deux méthodes dans la suite de cette section.

1.2.1 Sélection de modèle

La méthode de sélection de modèle par contraste pénalisé développée par Barron, Birgé et Massart (1999) (voir également Massart 2007) permet de définir, dans des contextes variés, des estimateurs adaptatifs. Nous nous intéresserons ici à un exemple particulier : celui de la régression à design aléatoire.

Un exemple illustratif : la régression à design aléatoire scalaire

Pour mieux comprendre cette méthode, nous laissons de côté pour un moment les données fonctionnelles et nous nous intéressons au modèle de régression à design aléatoire étudié, dans un cadre général, par Baraud (2002) :

$$Y = s(X) + \varepsilon,$$

où X est une variable aléatoire à valeurs dans $[0,1]$, $s : [0,1] \to \mathbb{R}$ une fonction inconnue à estimer et ε un terme de bruit supposé centré et indépendant de X. L'objectif est d'estimer s à partir d'un échantillon $\{(Y_i, X_i), i = 1, \dots, n\}$ de copies de (X, Y).

Soit $(\varphi_j)_{j\geq 1}$ une base de $\mathbb{L}^2([0,1])$ à partir de laquelle nous construisons une famille d'espaces d'approximation ou *modèles* $\{S_m := \text{Vect}\,\{\varphi_1, \dots, \varphi_{D_m}\}, m = 1, \dots, N_n\}$, avec N_n un entier positif et pour tout m, $\dim(S_m) = D_m$. Soit $\widehat{s}_m$ *l'estimateur des moindres carrés* sur S_m, c'est-à-dire l'élément de S_m minimisant le *contraste des moindres carrés*

$$\gamma_n : t \mapsto \frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2.$$

La qualité de l'estimateur $\widehat{s}_m$ dépend fortement de la dimension D_m qui peut être ainsi vue comme un paramètre de lissage. Lorsque D_m est trop petite, l'espace S_m dans lequel est défini l'estimateur est de petite dimension et tous les éléments de cet espace sont, en général, éloignés de la fonction à estimer : l'estimateur est *biaisé*. En revanche, lorsque D_m augmente, la *variance* de l'estimateur augmente.
L'objectif est donc de sélectionner l'estimateur de la famille $\{\widehat{s}_m, m = 1, \dots, N_n\}$ réalisant le meilleur compromis biais-variance, c'est-à-dire, pour le risque $\mathbb{L}^2$, $\widehat{s}_{m^*}$ avec

$$m^* \in \arg\min_{m=1,\dots,N_n} \mathbb{E}\left[\|s - \widehat{s}_m\|^2_{\mathbb{L}^2([0,1])}\right],$$

où $\|t\|^2_{\mathbb{L}^2([0,1])} := \int_0^1 t^2(u)du$.

L'estimateur $\widehat{s}_{m^*}$, que l'on appellera *oracle*, est le meilleur estimateur qu'il est possible de sélectionner, au sens du risque $\mathbb{L}^2$. Il est incalculable en pratique (car s est inconnu) et l'objectif est donc de définir un critère capable de sélectionner un estimateur $\widehat{s}_{\widehat{m}}$ ayant des

performances similaires à celles de l'oracle.

Dans un premier temps, regardons la décomposition suivante du risque $\mathbb{L}^2$, obtenue grâce au théorème de Pythagore :

$$\mathbb{E}\left[\|\widehat{s}_m - s\|^2_{\mathbb{L}^2([0,1])}\right] = \|s_m - s\|^2_{\mathbb{L}^2([0,1])} + \mathbb{E}\left[\|\widehat{s}_m - s_m\|^2_{\mathbb{L}^2([0,1])}\right],$$

avec s_m la projection orthogonale de s sur S_m.

Le premier terme représente l'erreur minimale que l'on commet en approchant s par un élément de S_m, il peut donc être vu comme un terme de biais. Lorsque s appartient à un certain sous-espace $\mathcal{F}_\alpha$ de $\mathbb{L}^2([0,1])$ composé de fonctions de régularité α [2] nous pouvons majorer ce terme de la façon suivante :

$$\|s - s_m\|^2_{\mathbb{L}^2([0,1])} \leq C D_m^{-2\alpha},$$

avec C une constante positive (qui pourra varier d'une ligne à l'autre dans la suite de ce chapitre).

Le second terme peut être vu comme un terme de variance. Sous certaines hypothèses portant sur la densité de la loi de X et sur la base $(\varphi_j)_{j\geq 1}$, vérifiées par exemple par des bases locales comme des bases d'ondelettes ou d'histogrammes, nous avons

$$\mathbb{E}\left[\|\widehat{s}_m - s_m\|^2_{\mathbb{L}^2([0,1])}\right] \leq C D_m \frac{\sigma^2}{n},$$

avec C une constante positive.

Ces deux résultats permettent de majorer le risque de l'estimateur par

$$\mathbb{E}[\|\widehat{s}_m - s\|^2_{\mathbb{L}^2([0,1])}] \leq C\left(D_m^{-2\alpha} + D_m \frac{\sigma^2}{n}\right).$$

La meilleure dimension possible est donc de l'ordre de $D_m \sim n^{1/(2\alpha+1)}$ et dépend du paramètre inconnu α.

L'objectif est de définir un critère de sélection de la dimension imitant le comportement biais-variance du risque de l'estimateur. La théorie nous permet de connaître l'ordre de grandeur du terme de variance. En revanche, le terme de biais dépend de quantités inconnues en pratique. L'idée de la sélection de modèle est d'estimer le biais en utilisant le contraste γ_n, ici le contraste des moindres carrés, appliqué à l'estimateur. La dimension est sélectionnée en minimisant le critère

$$\gamma_n(\widehat{s}_m) + \kappa\sigma^2 \frac{D_m}{n},$$

où $\kappa > 0$ est une constante universelle c'est-à-dire qu'elle ne dépend pas des paramètres du modèle. La calibration de ce paramètre κ est un problème important en pratique, mais bien documenté à présent dans les contextes de régression usuels, nous renvoyons à Birgé et Massart (2007) et Baudry, Maugis et Michel (2012). Remarquons que si on fixe $\kappa = 2$ le critère de sélection de la dimension est très proche du critère C_p de Mallows (1973). La quantité $\kappa\sigma^2 D_m/n$ est appelée *pénalité*.

2. Par exemple $\mathcal{F}_\alpha$ peut être une boule (pour la norme de Besov) de l'espace de Besov $\mathcal{B}_{\alpha,2,\infty}$, nous renvoyons à DeVore et Lorentz (1993) et Birgé et Massart (2000) pour des définitions et des énoncés précis.

L'estimateur sélectionné $\widehat{s}_{\widehat{m}}$ vérifie l'inégalité suivante (Baraud, 2002, Théorème 1.1) :

$$\mathbb{E}[\|\widehat{s}_{\widehat{m}} - s\|^2_{\mathbb{L}^2([0,1])}] \leq C \left(\min_{n=1,\dots,N_n} \left\{ \|s - s_m\|^2 + \kappa\sigma^2 \frac{D_m}{n} \right\} + \frac{1}{n} \right). \tag{1.6}$$

Le risque de l'estimateur sélectionné est donc équivalent, à une constante multiplicative près, au risque de l'oracle. L'inégalité (1.6) est appelée *inégalité-oracle*. Elle est non-asymptotique et permet notamment de montrer que l'estimateur $\widehat{s}_{\widehat{m}}$ a la même vitesse de convergence que l'estimateur oracle, c'est-à-dire $n^{-2\alpha/(2\alpha+1)}$.

Si $\mathcal{F}_\alpha$ vérifie également

$$\inf_{\widehat{s}} \sup_{s \in \mathcal{F}_\alpha} \mathbb{E}\left[\|\widehat{s} - s\|_{\mathbb{L}^2([0,1])}\right] \geq C n^{-2\alpha/(2\alpha+1)},$$

où l'infimum est pris sur tous les estimateurs de s que l'on peut calculer à partir de l'échantillon $\{(X_i, Y_i), i = 1, \dots, n\}$, on dit alors que $n^{-2\alpha/(2\alpha+1)}$ est la *vitesse optimale au sens du risque minimax* (que l'on appellera plus brièvement par la suite *vitesse minimax*). Comme l'estimateur $\widehat{s}_{\widehat{m}}$ atteint cette vitesse il est dit *adaptatif au sens minimax*.

Les techniques présentées ici sont très générales et peuvent s'appliquer à des contextes statistiques extrêmement variés : recherche de courbes principales dans un nuage de points (Fischer, 2013), données censurées (Brunel, Comte et Guilloux, 2013), prédiction de séries temporelles (Alquier et Wintenberger, 2012), estimation de graphe (Giraud, Huet et Verzelen, 2012),... De nombreux autres contrastes que les contrastes de type moindres carrés peuvent être considérés, par exemple des contrastes de type log-vraisemblance (Massart, 2007), dans ce cas-là la méthode de sélection de modèle par contraste pénalisé englobe le critère de sélection d'Akaike, pour un choix particulier de la pénalité.

Une des limitations de cette méthode est qu'elle s'applique uniquement lorsque la famille d'estimateurs $\{\widehat{s}_\lambda, \lambda \in \Lambda\}$ est obtenue par minimisation d'un contraste (moindres carrés, log-vraisemblance,...), ce qui n'est pas le cas, par exemple, des estimateurs à noyau.

1.2.2 Sélection de fenêtre

Estimation adaptative à noyau d'une densité

Récemment, Goldenshluger et Lepski (2011) ont proposé une méthode adaptative de sélection de fenêtre pour l'estimation de la densité d'une variable aléatoire multivariée. Le critère de sélection proposé est basé lui aussi sur une imitation de la décomposition biais-variance mais le biais est, cette fois-ci, estimé en introduisant un estimateur intermédiaire. Présentons brièvement une méthode inspirée de celle de Goldenshluger et Lepski (2011), dans un contexte plus simple que celui de l'article original : celui de l'estimation d'une densité d'une variable aléatoire univariée X à partir d'un échantillon $\{X_1, \dots, X_n\}$ composé de copies de X. Nous nous intéressons à l'estimateur de Parzen-Rosenblatt :

$$\widehat{s}_h(t) := \frac{1}{n} \sum_{i=1}^{n} K_h(t - X_i),$$

où $K_h(u) := (1/h)K(u/h)$ avec K un noyau c'est-à-dire une application de $\mathbb{R}$ dans $\mathbb{R}_+$ telle que $\int_{\mathbb{R}} K(u)du = 1$.

Si l'on s'intéresse au risque lié à la distance $\mathbb{L}^2(\mathbb{R})$, la décomposition biais-variance s'écrit ici

$$\mathbb{E}\left[\|\widehat{s}_h - s\|^2_{\mathbb{L}^2(\mathbb{R})}\right] = \|s - \mathbb{E}[\widehat{s}_h]\|^2_{\mathbb{L}^2(\mathbb{R})} + \mathbb{E}\left[\|\widehat{s}_h - \mathbb{E}[\widehat{s}_h]\|^2_{\mathbb{L}^2(\mathbb{R})}\right].$$

Nous pouvons voir facilement que

$$\mathbb{E}[\widehat{s}_h(t)] = \int_{\mathbb{R}} K_h(t-u)s(u)du = K_h * s(t) = \int_{\mathbb{R}} K(u)s(t-hu)du,$$

où $*$ désigne le produit de convolution. Comme $\int_{\mathbb{R}} K(u)du = 1$, le biais de l'estimateur $\widehat{s}_h$ s'écrit donc

$$\|\mathbb{E}[\widehat{s}_h] - s\|^2_{\mathbb{L}^2(\mathbb{R})} = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} K(u)(s(t-hu) - s(t))du\right)^2 dt.$$

Supposons que s est β-höldérienne avec $\beta < 1$, c'est-à-dire que $|s(u) - s(v)| \leq C_H|u-v|^\beta$ et que $\int_{\mathbb{R}} |u|^\beta K(u)du = M_K < +\infty$, alors

$$\|s - \mathbb{E}[\widehat{s}_h]\|^2_{\mathbb{L}^2(\mathbb{R})} \leq C_H^2 M_K^2 h^{2\beta}.$$

Encore une fois, l'ordre du biais est inconnu [3]. Pour la variance on obtient plus simplement

$$\mathbb{E}\left[\|\widehat{s}_h - \mathbb{E}[\widehat{s}_h]\|^2_{\mathbb{L}^2(\mathbb{R})}\right] \leq \frac{1}{nh}\int_{\mathbb{R}} K^2(u)du = \frac{\|K\|^2_{\mathbb{L}^2(\mathbb{R})}}{nh},$$

ce qui permet de définir un estimateur de la variance de la façon suivante

$$V(h) = \kappa\frac{\|K\|^2_{\mathbb{L}^2(\mathbb{R})}}{nh},$$

ici κ est une constante universelle. L'originalité de la méthode réside dans la façon d'estimer le biais par

$$A(h) = \sup_{h' \in \mathcal{H}_n} \left(\|K_h * \widehat{s}_{h'} - \widehat{s}_{h'}\|^2_{\mathbb{L}^2(\mathbb{R})} - V(h')\right)_+,$$

où $\mathcal{H}_n$ est la collection de fenêtres choisie. L'heuristique d'un tel choix pour l'estimateur du biais est la suivante : si $\widehat{s}_{h'}$ estime s alors $K_h * \widehat{s}_{h'}$ estime $\mathbb{E}[\widehat{s}_h] = K_h * s$ donc $\|K_h * \widehat{s}_{h'} - \widehat{s}_{h'}\|^2_{\mathbb{L}^2(\mathbb{R})}$ estime le biais de l'estimateur $\widehat{s}_h$, on retire un terme $V(h')$ de l'ordre de la variance pour tenir compte de la variabilité de l'estimateur $\widehat{s}_{h'}$. Le même calcul est répété pour toutes les valeurs de h' possibles.

Finalement la fenêtre est sélectionnée de la manière suivante

$$\widehat{h} := \arg\min_{h \in \mathcal{H}_n}\{A(h) + V(h)\}.$$

Ce critère est une version simplifiée du critère de Goldenshluger et Lepski (2011) qui

3. Pour des fonctions β-höldériennes avec $\beta \geq 1$, la majoration du biais est identique, à une constante multiplicative près, si le noyau K est suffisamment régulier (Tsybakov, 2009, proposition 1.2 et définitions 1.2 et 1.3).

s'écrit de manière plus générale. L'estimateur $\widehat{s}_{\widehat{h}}$ sélectionné vérifie une inégalité de type oracle pour le risque $\mathbb{L}^2$ et atteint la vitesse minimax sur de larges classes de fonctions. La méthode proposée s'applique également à des risques $\mathbb{L}^p$ généraux ainsi qu'à des variables à valeurs dans $\mathbb{R}^d$, nous renvoyons à l'article original pour plus de précisions.

Cette méthode a été adaptée à d'autres cadres statistiques que celui de l'estimation de la densité (voir par exemple Doumic et al. 2012 ; Bertin, Lacour et Rivoirard 2013 ; Comte et Lacour 2013). De plus, l'idée d'utiliser des estimateurs auxiliaires (c'est-à-dire ici $K_h * \widehat{s}_{h'}$) dans la définition du critère s'est révélée extrêmement féconde. Cette idée a tout d'abord été reprise pour des problèmes de sélection de la dimension d'estimateurs par projection (Comte et Johannes, 2012 ; Bertin, Lacour et Rivoirard, 2013 ; Chagny, 2013b). Elle se place ici comme une alternative à la méthode de sélection de la dimension par contraste pénalisé et nous expliquerons en détail dans la section 1.4.2 comment nous avons mis à profit ces idées pour la sélection de fenêtre dans un contexte où l'espérance de l'estimateur ne s'écrit pas comme un produit de convolution.

Elle a également été adaptée à la sélection de paramètres plus complexes, par exemple le paramètre $\theta \in \mathbb{R}^d$ du modèle single-index $Y = g(\theta x') + \varepsilon$ (Lepski et Serdyukova, 2014) ou des poids pour des estimateurs *linéaires* c'est-à-dire dont l'espérance dépend linéairement de la fonction à estimer (Goldenshluger et Lepski, 2013a).

1.3 Modèle linéaire fonctionnel

Les résultats du chapitre 2 se placent dans le cadre du modèle linéaire fonctionnel que nous présentons dans cette section.

1.3.1 Définition du modèle

Le modèle linéaire fonctionnel introduit pour la première fois par Ramsay et Dalzell (1991) et défini dans sa forme actuelle par Cardot, Ferraty et Sarda (1999) suppose qu'il existe une dépendance linéaire entre Y et sa covariable X. Plus précisément nous supposons que

$$Y = \langle \beta, X \rangle + \varepsilon, \tag{1.7}$$

où β est un élément de $\mathbb{H}$ appelé *fonction de pente* et ε une variable aléatoire centrée, indépendante de X, de variance finie σ^2. Nous appellerons ε le *terme de bruit*. Nous disposons d'un échantillon $\{(X_i, Y_i), i = 1, \dots, n\}$ suivant le modèle (1.7) et l'objectif est d'estimer la fonction β.

1.3.2 Problèmes inverses et identifiabilité

En multipliant les deux côtés de l'équation (1.7) par $X(s)$ et en prenant l'espérance, nous voyons que la fonction β est solution de l'équation

$$\Gamma\beta = \mathbb{E}[YX] =: g, \tag{1.8}$$

où Γ est l'opérateur de covariance de X défini par (1.3). L'opérateur Γ doit donc être inversé pour pouvoir retrouver β à partir de g. La fonction à estimer β est donc solution

d'un problème inverse.

Notons que le modèle linéaire fonctionnel est identifiable si et seulement si l'équation (1.8) admet une unique solution, c'est-à-dire lorsque $\mathrm{Ker}(\Gamma) = \{0\}$ ce qui revient à supposer que

$$\{j \geq 1,\ \lambda_j = 0\} = \emptyset, \tag{1.9}$$

où $(\lambda_j)_{j\geq 1}$ est la suite des fonctions propres de l'opérateur Γ.
En effet, montrons dans un premier temps que cette condition est nécessaire : supposons qu'il existe un entier J_0 tel que $\lambda_{J_0} = 0$. Soit ψ_{J_0} une fonction propre de Γ associée à la valeur propre λ_{J_0} nous avons $\Gamma\psi_{J_0} = 0$, or

$$\langle \Gamma\psi_{J_0}, \psi_{J_0}\rangle = \mathbb{E}[\langle X, \psi_{J_0}\rangle^2],$$

ce qui implique que $\langle X, \psi_{J_0}\rangle = 0$ p.s. Donc si β vérifie l'équation (1.7) alors $\beta + \psi_{J_0}$ vérifie aussi cette équation et le modèle n'est pas identifiable.
Montrons maintenant que la condition (1.9) est suffisante. Soient β et β' des éléments de $\mathbb{H}$ vérifiant l'équation du modèle (1.7), alors β et β' vérifient également l'équation (1.8). En particulier $\Gamma(\beta - \beta') = 0$, ce qui implique que $\beta - \beta' \in \mathrm{Ker}(\Gamma)$ et $\beta = \beta'$ dès que (1.9) est vérifiée.

Selon Hadamard (1902), un problème inverse doit vérifier les propriétés suivantes pour être *bien posé* : il existe une solution, cette solution est unique et stable dans le sens où elle dépend des données de manière continue. Or, sous la condition (1.9), l'opérateur Γ est injectif et il est bien possible de l'inverser en posant

$$\Gamma^{-1} : f \to \sum_{j\geq 1} \frac{\langle f, \psi_j\rangle}{\lambda_j}\psi_j,$$

mais cet inverse n'est pas un opérateur continu et n'est pas défini pour tout $f \in \mathbb{H}$[4]. En particulier, la condition de stabilité n'est pas vérifiée : le problème (1.8) est donc *mal posé*.

Considérons maintenant la version empirique de (1.8) :

$$\Gamma_n\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n} Y_i X_i =: \widehat{g}. \tag{1.10}$$

L'opérateur Γ_n étant de rang fini, il n'est pas inversible. Toutefois, il peut arriver que l'équation (1.10) admette une solution, lorsque $\widehat{g} \in \mathrm{Im}(\Gamma_n)$. Cependant, certaines valeurs propres de l'opérateur Γ_n étant très proches de 0, le problème inverse reste mal conditionné et la solution $\widehat{\beta}$ a une très grande variance (un exemple de réalisation de $\widehat{\beta}$ est donné dans la Figure 1.2 à titre illustratif).

Toutes les procédures d'estimation de la fonction β consistent à régulariser le problème inverse (1.10). Notre cadre présente donc des similarités avec les travaux réalisés dans la communauté statistique des problèmes inverses (voir Cavalier 2011 pour un récent état de l'art) qui consistent à estimer une fonction β en observant une image bruitée $Y = A\beta + \varepsilon$

4. Notons toutefois que l'opérateur Γ^{-1} est défini sur l'ensemble $\bigcup_{j\geq 1} \mathrm{Vect}\{\psi_1, \ldots, \psi_J\}$ qui est dense dans $\mathbb{H}$ et d'autre part qu'il peut-être obtenu comme limite ponctuelle d'opérateurs continus donc mesurables (une telle suite est donnée dans l'Annexe A, équation (A.2), p.160) ce qui implique qu'il est lui-même mesurable.

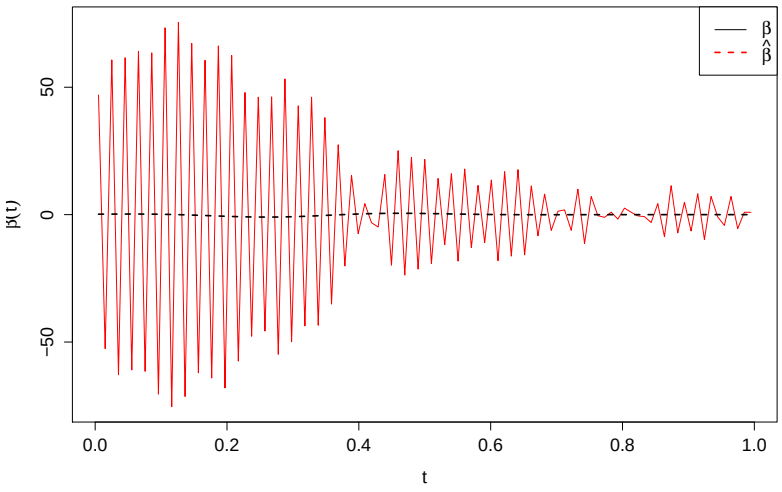


FIGURE 1.2 – Solution de $\Gamma_n\widehat{\beta} = \widehat{g}$, $n = 10\,000$, $\beta(t) = \exp(-(t-0.3)^2/0.05)\cos(4\pi t)$, X est le mouvement brownien sur $[0,1]$ et $\varepsilon \sim \sigma\mathcal{U}(-\sqrt{3},\sqrt{3})$ avec $\sigma^2 = 0.01$.

par un opérateur A. Cependant, les travaux réalisés dans ce contexte ne se transposent pas à notre cadre car l'opérateur A est soit supposé connu, soit les vecteurs propres de A sont supposés connus et ses valeurs propres observées de manière bruitée par un bruit indépendant de ε (Cavalier et Hengartner, 2005), or cette dernière hypothèse n'est pas réaliste dans le cadre du modèle linéaire fonctionnel.

L'estimateur B-spline proposé par Cardot, Ferraty et Sarda (2003) s'obtient par minimisation du critère des moindres carrés pénalisé

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle f, X_i\rangle)^2 + \rho\|f^{(m)}\|^2$$

sur l'espace des splines f définies sur $[0,1]$, de degré q à k intervalles égaux, avec $m < q$ un entier et $\rho > 0$ un paramètre de lissage. D'autres critères similaires ont été proposés, avec d'autres bases, comme la base de Fourier (Ramsay et Silverman, 2005 ; Li et Hsing, 2007) ou la base de l'ACP (Reiss et Ogden, 2007), ou d'autres pénalités (Crambes, Kneip et Sarda, 2009). Toutes ces procédures, qui mettent en jeu des pénalités de type ℓ^2, consistent à régulariser le problème inverse $\Gamma_n\widehat{\beta} = \widehat{g}$ avec la méthode de Tikhonov.

Une autre grande classe de méthodes d'estimation pour la fonction β est la classe des estimateurs par projection. Cela consiste à restreindre la matrice Γ_n à un sous-espace de $\mathbb{H}$ de dimension finie souvent assez petite. Cette base peut être fixe, par exemple la base de B-splines considérée par Ramsay et Dalzell (1991) ou la base de Fourier (Comte et Johannes, 2010) ou encore une base orthonormée générale (Cardot et Johannes, 2010). De nombreux travaux portent également sur des estimateurs par projection définis sur la base de l'ACP (voir section 1.1.1). L'estimateur des moindres carrés dans cette base s'écrit simplement de

la façon suivante

$$\widehat{\beta}_m^{(FPCR)} := \sum_{j=1}^{m} \frac{\langle \widehat{g}, \widehat{\psi}_j \rangle}{\widehat{\lambda}_j} \widehat{\psi}_j, \tag{1.11}$$

où m est un nombre entier, qui joue le même rôle ici que le paramètre de lissage ρ défini plus haut. Le comportement asymptotique de cet estimateur est bien connu : par exemple les convergences faibles en probabilité et p.s. ont été établies par Cardot, Ferraty et Sarda (1999) puis Cai et Hall (2006) ont montré des bornes supérieures et inférieures pour l'erreur de prédiction sur une courbe fixée x, enfin Hall et Horowitz (2007) ont étudié la vitesse de convergence et les propriétés d'optimalité de cet estimateur pour la norme $\mathbb{L}^2$. Des versions lissées ont également été introduites par Cardot, Ferraty et Sarda (2003).

1.3.3 Problème du choix de la dimension des estimateurs par projection

Toutes les procédures citées dans le paragraphe précédent reposent sur le choix d'au moins un paramètre de lissage : le paramètre ρ ou le nombre m de fonctions de base intervenant dans la définition de l'estimateur. Les choix optimaux théoriques de tels paramètres dépendent de quantités inconnues en pratique comme la régularité de β ou la vitesse de décroissance des valeurs propres $(\lambda_j)_{j\geq 1}$. Par exemple, si l'on considère l'estimateur par composantes principales défini par (1.11), le paramètre m minimisant l'erreur de prédiction $\langle \widehat{\beta}_K^{(FPCR)} - \beta, x\rangle^2$ sur une courbe fixée $x = \sum_{j\geq 1} x_j \psi_j$ telle que $x_j \leq C_1 j^{-\gamma}$, $\gamma > 1/2$ est de l'ordre de $n^{1/(a+2b-1)}$ où $a > 1$ est tel que $C_2^{-1} j^{-a} \leq \lambda_j \leq C_2 j^{-a}$ et $b > 1/2$ est tel que $\beta = \sum_{j\geq 1} \beta_j \psi_j$ avec $\beta_j \leq C_3 j^{-b}$ (on supposera ici $a + 1 > 2\gamma$).

En pratique, ces paramètres sont, la plupart du temps, sélectionnés par validation croisée (Cardot, Ferraty et Sarda, 2003 ; Hall et Hosseini-Nasab, 2006 ; Crambes, Kneip et Sarda, 2009). Récemment, des procédures d'estimation adaptatives de la fonction de pente β ont été proposées. Cai et Yuan (2012) ont établi, dans un contexte d'espace de Hilbert à noyau reproduisant, une procédure de sélection du paramètre de lissage ρ pour la méthode de régularisation développée par Ramsay et Silverman (2005), l'estimateur ainsi sélectionné atteint la vitesse de convergence optimale sur de larges classes de fonctions β. Une procédure adaptative de sélection de la dimension pour l'estimateur par projection défini par Cardot et Johannes (2010) a également été développée par Comte et Johannes (2010) avec des outils de sélection de modèle (voir section 1.2.1) puis généralisée en 2012 par les mêmes auteurs par une méthode inspirée de Goldenshluger et Lepski (2011). Leurs estimateurs vérifient une inégalité de type oracle et atteignent la vitesse de convergence minimax pour des risques $\mathbb{L}^2$ pondérés.

1.3.4 Contributions du chapitre 2

Dans ce contexte, le premier travail de thèse a consisté à définir une procédure de sélection de la dimension pour des estimateurs par projection dans le modèle linéaire fonctionnel (1.7). L'objectif était de choisir un bon espace d'approximation pour notre estimateur et nous nous sommes donc intéressés à la base de l'ACP. Dans le cas où les fonctions propres $(\psi_j)_{j\geq 1}$ de l'opérateur de covariance Γ sont connues, nous avons défini un estimateur des

moindres carrés $\widehat{\beta}_m^{(KB)}$ sur l'espace

$$S_m := \text{Vect}\{\psi_1, \ldots, \psi_{D_m}\},$$

de dimension D_m avec $(D_m)_{m\geq 1}$ une suite strictement croissante de nombres entiers naturels non nuls (des précisions supplémentaires sur le choix de D_m seront données dans le chapitre 2, section 2.2.2, p. 34). Ce premier travail a permis de se familiariser avec les problématiques d'estimation et de sélection de la dimension dans le modèle linéaire fonctionnel et d'identifier les points délicats pour le passage d'une base fixe à une base aléatoire. L'extension à l'estimateur des moindres carrés $\widehat{\beta}_m^{(FPCR)}$ sur l'espace aléatoire

$$\widehat{S}_m := \text{Vect}\{\widehat{\psi}_1, \ldots, \widehat{\psi}_{D_m}\}$$

engendré par les fonctions propres de l'opérateur de covariance empirique Γ_n, s'est faite dans un second temps.

Dans le chapitre 2, nous avons pris le parti de présenter les résultats pour les deux estimateurs $\hat{\beta}_m^{(KB)}$ et $\hat{\beta}_m^{(FPCR)}$ en parallèle, d'une part dans le souci d'éviter les répétitions, mais surtout pour donner un éclairage sur le passage délicat d'une base connue (fixe) à une base aléatoire. Nous espérons ainsi appuyer sur les points clés de notre contribution.

Nous fixons une dimension maximale D_{N_n} et sa version empirique $D_{\widehat{N}_n}$ (nous ne rentrons pas ici dans les détails, un peu techniques, des définitions de N_n et $\widehat{N}_n$ qui peuvent être trouvés dans la section 2.2.2, p.34). Nous sélectionnons ensuite un estimateur dans la famille $\left\{\widehat{\beta}_m^{(KB)}, m = 1, \ldots, \widehat{N}_n\right\}$ ou la famille $\left\{\widehat{\beta}_m^{(FPCR)}, m = 1, \ldots, \widehat{N}_n\right\}$ par minimisation du critère pénalisé

$$\text{crit}(m) := \gamma_n(\widehat{\beta}_m) + \kappa\widehat{\sigma}_m^2 \frac{D_m}{n}, \tag{1.12}$$

où $\widehat{\sigma}_m^2$ est un estimateur de la variance du bruit σ^2,

$$\widehat{\sigma}_m^2 := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \langle\widehat{\beta}_m, X_i\rangle\right)^2,$$

et γ_n le contraste des moindres carrés

$$\gamma_n : f \mapsto \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \langle f, X_i\rangle\right)^2.$$

La forme usuelle de la pénalité en régression est $\kappa\sigma^2\frac{D_m}{n}$ (voir par exemple Baraud 2002 ; Massart 2007). Cependant cette pénalité dépend de la variance du bruit qui est, en pratique, souvent inconnue, d'où l'introduction de l'estimateur plug-in $\widehat{\sigma}_m^2$. Remarquons que Baraud, Giraud et Huet (2009) ont également introduit un critère *multiplicatif* très similaire pour sélectionner la dimension d'estimateurs par projection dans le modèle de régression gaussien, lorsque la variance du bruit est inconnue.

Pour évaluer les performances de l'estimateur obtenu, nous nous intéressons à l'erreur

de prédiction, c'est-à-dire à la quantité

$$\mathbb{E}\left[\left(\widehat{Y}_{n+1} - \mathbb{E}[Y_{n+1}|X_{n+1}]\right)^2 |(X_1,Y_1),...,(X_n,Y_n)\right] = \|\Gamma^{1/2}(\widehat{\beta}-\beta)\|^2 =: \|\widehat{\beta}-\beta\|^2_\Gamma,$$

où (X_{n+1},Y_{n+1}) est une copie de (X,Y) indépendante de l'échantillon et $\widehat{Y}_{n+1} := \langle\widehat{\beta}, X_{n+1}\rangle$.

Nous montrons dans un premier temps (Proposition 1, p. 36) que les deux estimateurs obtenus vérifient une inégalité de type oracle pour la version empirique $\|\widehat{\beta}-\beta\|^2_{\Gamma_n} = \|\Gamma_n^{1/2}(\widehat{\beta}-\beta)\|^2$ de l'erreur de prédiction. Cette inégalité de type oracle s'écrit

$$\mathbb{E}\left[\|\widehat{\beta}^{(KB)}_{\widehat{m}} - \beta\|^2_{\Gamma_n}\right] \leq C\left(\inf_{m=1,\dots,N_n}\left\{\mathbb{E}\left[\|\beta - \Pi_m\beta\|^2_{\Gamma_n}\right] + \mathrm{pen}(m)\right\} + \frac{\sigma^2+\|\beta\|^2}{n}\right), \quad (1.13)$$

pour l'estimateur $\widehat{\beta}^{(KB)}_m$ (ici $\mathcal{M}_n$ est la collection d'indices dans laquelle est choisi $\widehat{m}$ et Π_m l'opérateur de projection sur S_m) et

$$\mathbb{E}\left[\|\widehat{\beta}^{(FPCR)}_{\widehat{m}} - \beta\|^2_{\Gamma_n}\right] \leq C\left(\inf_{m=1,\dots,N_n}\left\{\mathbb{E}\left[\|\beta - \widehat{\Pi}_m\beta\|^2_{\Gamma_n}\right] + \mathrm{pen}(m)\right\} + \frac{\sigma^2+\|\beta\|^2}{n}\right), \quad (1.14)$$

pour l'estimateur $\widehat{\beta}^{(FPCR)}_{\widehat{m}}$ (avec $\widehat{\Pi}_m$ l'opérateur de projection sur l'ensemble aléatoire $\widehat{S}_m$).

En montrant que, pour tout $\rho_0 \in]0,1[$, pour tout $m = 1,\dots,N_n$, $\sup_{f\in S_m}\frac{\|f\|^2_{\Gamma_n}}{\|f\|^2_\Gamma} < \rho_0$ avec grande probabilité, nous obtenons une inégalité de type oracle pour l'estimateur $\widehat{\beta}^{(KB)}_{\widehat{m}}$ (Équation (2.10) du théorème 1, p.37) :

$$\mathbb{E}\left[\|\widehat{\beta}^{(KB)}_{\widehat{m}} - \beta\|^2_\Gamma\right] \leq C_1\left(\min_{m\in\mathcal{M}_n}\left(\|\beta-\Pi_m\beta\|^2_\Gamma + \mathrm{pen}(m)\right)\right) + \frac{C_2}{n}. \quad (1.15)$$

L'extension de ces résultats à l'estimateur $\widehat{\beta}^{(FPCR)}$ défini sur des bases aléatoires pose certains problèmes techniques : tout d'abord le terme $\sup_{f\in\widehat{S}_m}\frac{\|f\|^2_{\Gamma_n}}{\|f\|^2_\Gamma}$ est plus difficile à contrôler que $\sup_{f\in S_m}\frac{\|f\|^2_{\Gamma_n}}{\|f\|^2_\Gamma}$. En effet, $\sup_{f\in S_m}\frac{\|f\|^2_{\Gamma_n}}{\|f\|^2_\Gamma}$ est la plus grande valeur propre de la matrice $\left(\langle\Gamma_n\psi_j,\psi_k\rangle/\sqrt{\lambda_j\lambda_k}\right)_{1\leq j,k\leq D_m}$ dont les coefficients sont des moyennes de variables aléatoires indépendantes, ce qui permet de borner cette quantité directement avec une inégalité de type Bernstein (Lemme 22, p.180). En revanche, l'aléa présent dans le terme $\sup_{f\in\widehat{S}_m}\frac{\|f\|^2_{\Gamma_n}}{\|f\|^2_\Gamma}$, qui dépend du spectre de la matrice $\left(\langle\Gamma\widehat{\psi}_j,\widehat{\psi}_k\rangle/\sqrt{\widehat{\lambda}_j\widehat{\lambda}_k}\right)_{1\leq j,k\leq D_m}$, est plus difficile à contrôler. D'autre part, le terme de biais de l'inégalité (1.15) s'obtient aisément en constatant que

$$\mathbb{E}\left[\|\beta-\Pi_m\beta\|^2_{\Gamma_n}\right] = \|\beta-\Pi_m\beta\|^2_\Gamma,$$

mais cette égalité n'est plus vraie si l'on remplace le projecteur Π_m par sa version empirique $\widehat{\Pi}_m$. L'inégalité suivante (Inégalité (2.11) du théorème 1, p.37)

$$\mathbb{E}[\|\widehat{\beta}^{(FPCR)}_{\widehat{m}} - \beta\|^2_\Gamma] \leq C_1'\left(\min_{m\in\mathcal{M}_n}\left(\mathbb{E}[\|\beta-\widehat{\Pi}_m\beta\|^2_\Gamma] + \mathrm{pen}(m)\right)\right) + \frac{C_2'}{n} \quad (1.16)$$

n'a donc pu être prouvée qu'avec l'aide des outils puissants de la théorie de la perturbation présentés en Annexe A et au prix d'hypothèses supplémentaires, notamment sur la fonction à estimer.

À l'aide des deux inégalités oracles (1.15) et (1.16) nous montrons (Théorème 2, p.38) une majoration de la vitesse de convergence de l'erreur de prédiction des deux estimateurs $\widehat{\beta}_{\widehat{m}}^{(KB)}$ et $\widehat{\beta}_{\widehat{m}}^{(FPCR)}$. La vitesse ainsi obtenue coïncide avec la borne inférieure déterminée par Cardot et Johannes (2010) ce qui implique que nos deux estimateurs sont optimaux au sens minimax.

Nous nous sommes ensuite intéressés aux propriétés numériques de ces deux estimateurs. Dans la section 2.4.3 p.42, plusieurs méthodes (calibration par simulations, heuristique de pente et détection de saut de dimension) sont comparées pour la calibration de la constante κ apparaissant dans la définition du critère (1.12). Les deux estimateurs sont ensuite étudiés dans différents contextes dans la section 2.4.5 p.47. Nous nous focalisons ensuite sur l'estimateur $\widehat{\beta}_{\widehat{m}}^{(FPCR)}$. Dans la section 2.4.6, notre méthode de sélection de la dimension est comparée à deux critères de validation croisée fréquemment utilisés dans la littérature. Les résultats obtenus nous permettent d'affirmer que notre méthode présente de sérieux avantages, tant du point de vue du temps de calcul, que du point de vue de la stabilité des estimateurs obtenus.

1.4 Estimation des fonctionnelles de régression

Nous nous intéresserons ici à des modèles statistiques où très peu d'hypothèses sont faites sur la relation entre Y et X. Dans un premier temps, nous allons nous concentrer sur le modèle de régression non-paramétrique fonctionnel sur lequel portent l'essentiel des travaux théoriques existant dans la littérature sur des modèles non-paramétriques avec covariable fonctionnelle. Ce modèle s'écrit

$$Y = m(X) + \varepsilon$$

où m est une fonction de $\mathbb{H} \to \mathbb{R}$ et ε un terme de bruit supposé indépendant de X. L'objectif est d'estimer la fonction m à partir d'un échantillon $\{(X_i, Y_i), i = 1, \ldots, n\}$ de copies de (X, Y).

Dans un second temps, nous nous intéresserons à l'estimation de la fonction de répartition de Y sachant $X = x$

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

qui est l'objet des travaux présentés dans le chapitre 3.

1.4.1 Fléau de la dimension

Les estimateurs connus de la fonction de régression m sont des estimateurs de type Nadaraya-Watson :

$$\widehat{m}_h(x) := \frac{\sum_{i=1}^n K_h(d(X_i, x))Y_i}{\sum_{i=1}^n K_h(d(X_i, x))}, \tag{1.17}$$

où K est un noyau, d une semi-norme sur $\mathbb{H}$ et $K_h(t) := (1/h)K(t/h)$.

Le premier travail portant sur la fonction de régression dans le cadre fonctionnel est celui de Ferraty et Vieu (2002). La semi-norme d est ici une semi-norme basée sur les dérivées

$d_m(f,g)^2 = \int \left(f^{(m)}(t) - g^{(m)}(t)\right)^2 dt$ avec m un entier et $f^{(m)}$ la dérivée m-ème de f. La fenêtre h et l'entier m sont sélectionnés par validation croisée. La convergence ponctuelle de cet estimateur est prouvée. Cette procédure a été ensuite généralisée à des données dépendantes et avec des semi-normes générales par Ferraty et Vieu (2004) et Benhenni, Hedli-Griche et Rachdi (2010) lorsque le design est fixe, puis lorsque la variable Y est également fonctionnelle (Lian, 2012). Une version robuste a été définie par Crambes, Delsol et Laksaci (2008).

Cet estimateur est convergent pour la convergence uniforme presque complète (Ferraty, Laksaci, Tadj et al., 2010). En revanche, la vitesse de convergence de ces estimateurs est souvent assez lente.

En effet, la variance de l'estimateur (1.17) est fonction de l'inverse de la probabilité de petite boule $\varphi^x(h) = \mathbb{P}\left(d(X,x) \leq h\right)$, typiquement de l'ordre de $\sqrt{\ln n/(n\varphi^x(h))}$ pour un risque non quadratique (voir par exemple Ferraty, Laksaci et Vieu 2006, Théorème 3.1). C'est donc le comportement asymptotique de $\varphi^x(h)$ quand $h \to 0$ qui détermine la vitesse de convergence de l'estimateur $\widehat{m}_h$.

Lorsque $\mathbb{H}$ est de dimension finie, où lorsque le rang de l'opérateur Γ est fini (c'est-à-dire que X peut s'écrire presque sûrement comme combinaison linéaire aléatoire d'un nombre fini d'éléments de $\mathbb{H}$) la probabilité de petite boule décroît typiquement à vitesse polynomiale vers 0 c'est-à-dire $\varphi(h) \sim_{h\to 0} Ch^p$ avec $p = \mathrm{rg}(\Gamma)$. En revanche, lorsque $\mathrm{rg}(\Gamma) = +\infty$, $h^{-p}\varphi(h) \to_{h\to 0} 0$ pour tout entier $p > 0$ (Mas, 2012, Corollaire 1). À titre illustratif, lorsque X est le pont brownien sur $[0,1]$, $\varphi(h) \sim_{h\to 0} c\exp\left(-\frac{1}{8h^2}\right)$ (Li et Shao, 2001, Proposition 6).

Autrement dit, plus X évolue dans un espace de dimension élevée et moins l'on trouvera d'observations autour de 0. Cela implique que la variance de l'estimateur $\widehat{m}_h$ diverge très vite vers $+\infty$ lorsque $h \to 0$. Comme le biais est en général polynomial, par exemple de l'ordre de h^β, il ne décroît pas assez vite vers 0 pour compenser suffisamment le terme de variance, ce qui explique la vitesse de convergence très lente des estimateurs à noyau.

Toutefois, le problème n'est pas lié au choix de l'estimateur mais à la complexité du problème posé. En effet, les bornes inférieures prouvées par Mas (2012) indiquent qu'aucun estimateur de la fonction de régression (à noyau ou non) ne peut atteindre une vitesse de convergence polynomiale lorsque X évolue dans un espace de dimension infinie.

1.4.2 Contributions du chapitre 3

Dans le chapitre 3, en collaboration avec Gaëlle Chagny (LMRS), nous proposons une procédure d'estimation adaptative pour la fonction de répartition $F^X : y \mapsto \mathbb{P}(Y \leq y|X)$ d'une variable réelle Y conditionnellement à une variable fonctionnelle X. Nous nous intéressons à l'estimateur de type Nadaraya-Watson proposé par Ferraty, Laksaci et Vieu (2006) et Ferraty, Laksaci, Tadj et al. (2010) défini par

$$\widehat{F}^x_{h,d}(y) := \frac{\sum_{i=1}^n \mathbf{1}_{\{Y_i \leq y\}} K_h(d(X_i,x))}{\sum_{i=1}^n K_h(d(X_i,x))},$$

où $K_h(u) = (1/h)K(u/h)$.

Pour contourner le fléau de la dimension, de nombreux auteurs (voir par exemple Ferraty, Laksaci et Vieu 2006 ; Delsol 2010 ; Geenens 2011) ont proposé de prendre une semi-norme

basée sur les projections

$$d_p^2(x,x') := \sum_{j=1}^{p} \langle x - x', e_j \rangle^2,$$

avec $(e_j)_{j\geq 1}$ une base hilbertienne de $\mathbb{H}$. Notons que cela revient à projeter X dans l'espace $\text{Vect}\{e_1,\ldots,e_p\}$. Pour plus de simplicité, nous noterons $\widehat{F}_h^x = \widehat{F}_{h,d}^x$ lorsque $d(x,x') = \|x-x'\|$ et $\widehat{F}_{h,p}^x = \widehat{F}_{h,d_p}^x$.

L'objectif initial de ce travail était de proposer une méthode de sélection simultanée de la fenêtre h et de la dimension de l'espace d'approximation p et d'associer la méthode de Goldenshluger et Lepski (2011) aux méthodes de réduction de la dimension en statistique pour données fonctionnelles.

Dans une première étape nous avons cherché à mieux comprendre le comportement du risque de l'estimateur en fonction de p et h. Nous nous sommes intéressées à un risque intégré

$$\mathbb{E}\left[\|\widehat{F}_{h,p}^{X'} - F^{X'}\|_D^2 \mathbf{1}_B(X')\right] = \mathbb{E}\left[\int_B \left(\int_D \left(\widehat{F}_{h,p}^x(y) - F^x(y)\right)^2 dy\right) d\mathbb{P}_X(x)\right],$$

avec X' une copie de X indépendante de l'échantillon et B un sous-ensemble borné de $\mathbb{H}$. Le premier résultat (Théorème 3, p.79), obtenu à "$p = +\infty$", c'est-à-dire avec $d(x,x') = \|x-x'\|$ exhibe une décomposition biais-variance de la forme suivante

$$\mathbb{E}\left[\|\widehat{F}_h^{X'} - F^{X'}\|_D^2 \mathbf{1}_B(X')\right] \leq C\left(h^{2\beta} + \frac{1}{n\varphi(h)}\right),$$

où β est l'indice de régularité de la fonction $x \mapsto F^x$ (on suppose ici que cette fonction est β-höldérienne). Ce résultat est cohérent avec ceux obtenus par Ferraty, Laksaci et Vieu (2006, Théorème 3.1) en terme de convergence presque complète.

Concernant l'estimateur calculé à partir de la semi-norme d_p, nous obtenons la majoration suivante (Proposition 2, p.87) :

$$\mathbb{E}\left[\|\widehat{F}_{h,p}^{X'} - F^{X'}\|_D^2 \mathbf{1}_B(X')\right] \leq C\left(h^{2\beta} + \left(\sum_{j>p} \sigma_j^2\right)^{\beta} + \frac{1}{n\varphi_p(h)}\right), \tag{1.18}$$

avec $\varphi_p(h) = \mathbb{P}(d_p(X,0) \leq h)$ et $\sigma_j^2 = \text{Var}(\langle X, e_j\rangle)$, pour tout $j \geq 1$. Nous avons ici un terme de biais supplémentaire $\left(\sum_{j>p}\sigma_j^2\right)^{\beta}$ par rapport aux résultats de Ferraty, Laksaci et Vieu (2006, Théorème 3.1) qui est dû à la perte d'information lors de l'étape de projection. Cette différence dans les résultats est liée à une différence dans les hypothèses de régularité imposées à la fonction $x \mapsto F^x$. Nous supposons (hypothèse H_F, p.79) que pour tous $x, x' \in \mathbb{H}$

$$\|F^x - F^{x'}\|_D \leq C\|x - x'\|^{\beta},$$

avec C une constante positive ; tandis que l'hypothèse (H2) de Ferraty, Laksaci et Vieu (2006) revient à supposer que, pour tout y dans un sous-ensemble de $\mathbb{R}$ et pour tous x, x'

dans un sous-ensemble de $\mathbb{H}$,

$$|F^x(y) - F^{x'}(y)| \leq C d_p(x, x')^\beta.$$

Or, cette hypothèse (H2) de Ferraty, Laksaci et Vieu (2006) fait intervenir le paramètre p d'estimation et ne permet donc pas de comparer les propriétés des estimateurs $\widehat{F}^x_{h,p}$ lorsque p varie. D'autre part, remarquons que cette hypothèse (H2) est, de manière implicite, une hypothèse de réduction de la dimension puisque, si l'on prend $x \neq x'$ avec $\langle x, e_j\rangle = \langle x', e_j\rangle$ pour tout $j \leq p$, alors $F^x = F^{x'}$; F^x ne dépend donc que de la projection de x dans un espace de dimension finie connue du statisticien (puisque le paramètre p ainsi que la base $(e_j)_{j\geq 1}$ interviennent dans la définition de l'estimateur). Une hypothèse de type H_F nous paraissait donc plus réaliste.

Une fois ces deux majorations obtenues, nous nous sommes concentrées dans une seconde étape sur la détermination d'une procédure de sélection de la fenêtre h de l'estimateur $\widehat{F}^x_h$. L'objectif était d'adapter à notre cadre la méthode de Goldenshluger et Lepski (2011). L'absence de définition de la densité d'une variable aléatoire fonctionnelle ne permet pas d'écrire le biais d'un tel estimateur sous la forme d'un produit de convolution. Une problématique très similaire se pose dans des contextes de sélection de modèle lorsque l'on cherche à sélectionner un estimateur dans une famille $\{\widehat{s}_m, m = 1, \ldots, N_n\}$ d'estimateurs par projection sur des modèles $(S_m)_{m=1,\ldots,N_n}$. Lorsque les modèles sont emboîtés (c'est-à-dire que $S_m \subset S_{m'}$ lorsque $m \leq m'$), Comte et Johannes (2012) et Chagny (2013b) ont proposé des procédures adaptatives de sélection de la dimension de l'espace d'approximation via une adaptation de la méthode de Goldenshluger et Lepski (2011) mettant en jeu des estimateurs auxiliaires de la forme $\widehat{s}_{m\wedge m'}$. Par analogie, nous nous sommes inspirées de ces travaux pour définir un estimateur auxiliaire de la fonction de répartition conditionnelle de la forme $\widehat{F}^x_{h\vee h'}$. La fenêtre $\widehat{h}$ est donc sélectionnée par minimisation du critère

$$\mathrm{crit}(h) = \widehat{A}(h) + \widehat{V}(h),$$

où

$$\widehat{V}(h) = \begin{cases} \kappa \dfrac{\ln(n)}{n\widehat{\varphi}(h)} \text{ si } \widehat{\varphi}(h) \neq 0 \\ +\infty \text{ sinon,} \end{cases} \qquad \text{avec } \widehat{\varphi}(h) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{\|X_i\|\leq h\}},$$

est un estimateur de la variance et

$$\widehat{A}(h) = \max_{h'\in\mathcal{H}_n} \left(\left\| \widehat{F}^{X'}_{h'} - \widehat{F}^{X'}_{h\vee h'} \right\|^2_D - \widehat{V}(h') \right)_+$$

estime le biais.

L'estimateur $\widehat{F}^x_{\widehat{h}}$ ainsi obtenu atteint la même vitesse de convergence que l'*oracle* (Théorème 4, p.82 et tableau 3.1 lignes a) et b)) à une perte logarithmique près.

Nous nous sommes ensuite concentrées sur l'estimateur $\widehat{F}^x_{h,p}$. Une première étape a été de déterminer les vitesses de convergence de cet estimateur. À cette étape nous fûmes assez surprises de constater que la majoration (1.18) ne permettait, au mieux, c'est-à-dire en choisissant bien les paramètres d'estimation (h, p) ainsi que la base $(e_j)_{j\geq 1}$, que d'atteindre la vitesse de convergence de l'estimateur $\widehat{F}^x_h$, ce qui est en contradiction avec les arguments

présentés par Geenens (2011) et Delsol (2010).

Les bornes inférieures que nous avons ensuite prouvées (Théorème 6, p.85, voir aussi la ligne c) du tableau 3.1) ont permis d'établir que les vitesses de convergence de l'oracle $\widehat{F}^x_{h^*}$ étaient en réalité optimales au sens minimax et donc que l'estimateur $\widehat{F}^x_{h,p}$ ne pouvait pas converger à une vitesse supérieure. Grâce à ce résultat, nous pouvons également affirmer que notre estimateur $\widehat{F}^x_{\widehat{h}}$ atteint cette vitesse minimax soit de manière exacte, il est dans ce cas-là adaptatif, soit à une perte logarithmique près, suivant le comportement de la probabilité de petite boule $\varphi(h)$ au voisinage de 0.

Notons que la méthode de sélection de la fenêtre ainsi que les résultats concernant l'estimateur $\widehat{F}^x_{\widehat{h}}$ sont également valables lorsque $\mathbb{H}$ est de dimension finie. Par exemple, lorsque $\mathbb{H} = \mathbb{R}^d$, nous retrouvons la vitesse minimax de l'ordre de $n^{-2\beta/(2\beta+d)}$.

Des résultats similaires sont obtenus avec un risque ponctuel

$$\mathbb{E}\left[\|\widehat{F}^{x_0}_h - F^{x_0}\|^2_D\right] = \mathbb{E}\left[\int_D \left(\widehat{F}^{x_0}_h(y) - F^{x_0}(y)\right)^2 dy\right],$$

où x_0 est un élément de $\mathbb{H}$.

1.4.3 Perspectives : modèles avec contrainte structurelle

Les arguments présentés dans la section précédente tendent à mener vers la conclusion que le modèle de régression non-paramétrique est trop général pour pouvoir définir des estimateurs suffisamment efficaces. D'un autre côté, le modèle linéaire, présenté dans la section 1.3, peut être trop restrictif pour certaines applications. D'où l'intérêt de se placer dans des modèles plus généraux qui toutefois ont l'avantage d'échapper au fléau de la dimension comme le modèle linéaire généralisé ou le modèle à direction révélatrice unique (connu également sous le nom de single-index). À notre connaissance, les travaux permettant de définir des estimateurs adaptatifs pour ces modèles sont encore inexistants, ce qui fournit des perspectives logiques pour la suite de ce travail que nous détaillerons à la fin de ce manuscrit.

Modèle linéaire généralisé

Le modèle le plus simple dans ce cadre est le modèle linéaire généralisé qui s'écrit de manière classique de la façon suivante

$$g(\mathbb{E}[Y|X]) = \alpha + \langle \beta, X \rangle,$$

où $g : \mathbb{R} \to \mathbb{R}$ est une fonction connue, monotone et inversible, la *fonction de lien*, et $\alpha \in \mathbb{R}$ et $\beta \in \mathbb{H}$ sont des paramètres inconnus à estimer. Une autre manière d'écrire ce modèle est la suivante

$$Y = g^{-1}\left(\alpha + \langle \beta, X \rangle\right) + \varepsilon,$$

où ε est un terme de bruit. Cette définition regroupe plusieurs types de modèles comme le modèle de régression logistique ($g(t) = \ln(t/(1-t))$) ou le modèle linéaire fonctionnel ($g(t) = t$). Un des premiers travaux sur le modèle linéaire généralisé fonctionnel est celui de James (2002) qui propose une procédure basée sur l'algorithme EM. Müller et Stadtmüller

(2005) ont ensuite proposé une procédure d'estimation de la fonction β par projection dans une base orthonormée de $\mathbb{L}^2$ et minimisation d'une fonction de score. Cardot et Sarda (2005) ont généralisé la procédure d'estimation préalablement proposée dans le contexte du modèle linéaire fonctionnel (Cardot, Ferraty et Sarda, 2003) et prouvé des vitesses de convergence pour le risque $\mathbb{L}^2$. Un certain nombre de travaux a porté uniquement sur le modèle logistique, la fonction β étant estimée par exemple par une approche de réduction de la dimension basée sur l'ACP (Escabias, Aguilera et Valderrama, 2004, 2005 ; Aguilera, Escabias et Valderrama, 2006, 2008) éventuellement associée à une étape de lissage (Aguilera, Aguilera-Morillo et al., 2011 ; Aguilera-Morillo et al., 2013). Des extensions de ce modèle à un nombre plus important de covariables ont été considérées, par exemple par James et Silverman (2005) et Li, Wang et Carroll (2010).

Modèle single-index

La définition du modèle single-index est tout à fait analogue à celle du modèle linéaire généralisé excepté que la fonction de lien g est ici également inconnue. Ce type de modèle a été développé dans des contextes où la covariable est multivariée (Ichimura, 1993 ; Newey et Stoker, 1993 ; Delecroix et Hristache, 1999 ; Bouaziz, 2010), comme intermédiaire entre les modèles non-paramétriques sujets au fléau de la dimension et les modèles linéaires généralisés qui peuvent être trop restrictifs pour certaines applications. Ce modèle à encore été très peu étudié lorsque la covariable est fonctionnelle : à notre connaissance, seuls deux travaux (Ait-Saïdi et al., 2008 ; Chen, Hall et Müller, 2011) se placent dans ce contexte. En particulier, Chen, Hall et Müller (2011) ont montré que l'erreur quadratique moyenne de leur estimateur atteint une vitesse de convergence polynomiale. Bien que l'ordre exact de la vitesse soit inconnu, ce résultat est encourageant pour continuer l'étude des modèles single-index.

Une généralisation du modèle single-index consisterait à supposer que la dépendance de Y par rapport à X dépend non pas de la projection de X sur une seule droite $\langle \beta, X \rangle$ de l'espace $\mathbb{H}$ mais sur un sous-espace de dimension finie de $\mathbb{H}$. Une première manière de considérer cette relation est de supposer l'existence de m éléments $\beta_1, \ldots, \beta_m$ de $\mathbb{H}$ et d'une fonction $g : \mathbb{R}^m \to \mathbb{R}$ inconnue tels que

$$Y = g(\langle \beta_1, X \rangle, \ldots, \langle \beta_m, X \rangle) + \varepsilon.$$

Ce modèle, bien qu'évoqué dans l'introduction de l'article de Chen, Hall et Müller (2011), n'a pas été étudié, probablement en raison de sa complexité. En revanche, des versions additives de ces modèles, qui peuvent être vues comme des généralisations fonctionnelles du modèle de *projection pursuit* (Friedman et Stuetzle, 1981),

$$Y = \sum_{j=1}^{m} g_j \left(\langle \beta_j, X \rangle \right) + \varepsilon$$

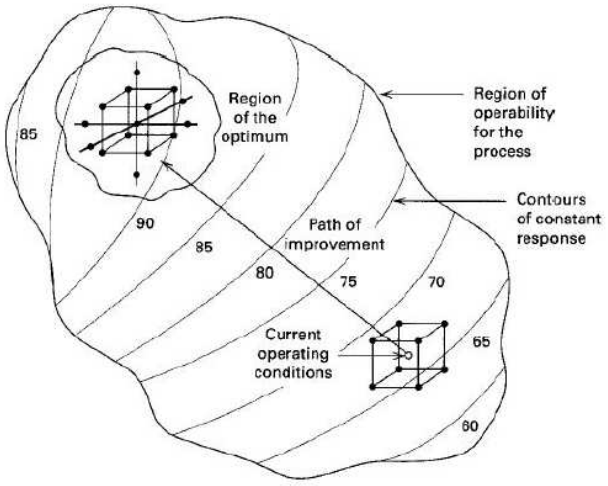
ont été considérées par exemple par Chen, Hall et Müller (2011) sous le nom de *multiple-index model* ou par James et Silverman (2005), dans un contexte paramétrique, sous le nom de *functional adaptive model*. Cependant, il n'existe pas, à ce jour, de résultat théorique sur ces modèles lorsque la covariable est fonctionnelle.

1.5 Surfaces de réponse

La méthode des surfaces de réponse a été introduite pour la première fois par Box et Wilson (1951) dans l'objectif de déterminer des conditions optimales pour des expérimentations en chimie : minimiser le coût de l'expérimentation ou maximiser la pureté du produit obtenu en trouvant la bonne combinaison de facteurs (température, pression, proportion des réactifs,...). Son but est donc d'étudier le lien entre une ou plusieurs variables explicatives $(x_1, \ldots, x_d) \in \mathbb{R}^d$ et une variable réponse $Y \in \mathbb{R}$ et en particulier de trouver les valeurs des variables explicatives pour lesquelles la variable réponse est minimale (ou maximale). Cette méthode a été et est toujours très largement utilisée dans l'industrie.

1.5.1 Principe de la méthode : optimisation séquentielle

Le principe de la méthode est de trouver les conditions d'expérimentation optimales en réalisant un nombre restreint d'expériences. Le point de départ est une certaine région de l'espace $\mathbb{R}^d$ (les conditions d'expérimentation actuelles que l'on cherche à optimiser par exemple) dans laquelle on réalise une série d'expériences, en suivant un *plan d'expériences* choisi à l'avance. Les observations ainsi obtenues nous permettent d'avoir une certaine idée de la forme de la surface $y = m(x_1, \ldots, x_n) := E[Y|x_1, \ldots, x_n]$ dans cette région. On utilise cette connaissance pour estimer la direction de plus forte descente (ou de plus forte croissance suivant que l'on cherche à maximiser ou minimiser m) de cette surface. Le long de cette direction, une série d'expériences est effectuée jusqu'à un point de $\mathbb{R}^d$ où la réponse est considérée optimale. Une série d'expériences est ensuite réalisée autour de ce point permettant soit de définir une autre direction de descente, soit d'affiner la position de l'optimum. La figure suivante, extraite de Montgomery (2009), illustre le déroulement de la méthode.



Le principe est resté globalement inchangé depuis les travaux de Box et Wilson (1951). Les principales améliorations ont été faites sur deux aspects : la modélisation de la surface dans la région considérée et le choix du *plan d'expériences*, c'est-à-dire des points où l'on réalise les expériences.

La manière classique de modéliser la forme de la surface dans la région d'intérêt est de considérer des modèles de régression polynomiaux, souvent de degré 1

$$Y = \alpha + \sum_{j=1}^{d} \beta_j x_j + \varepsilon,$$

avec $\alpha \in \mathbb{R}$, $(\beta_1, \ldots, \beta_j) \in \mathbb{R}^d$ et $\varepsilon \sim \mathcal{N}(0, 1)$, ou de degré 2

$$Y = \alpha + \sum_{j=1}^{d} \beta_j x_j + \sum_{j,k=1}^{d} \beta_{j,k} x_j x_k + \varepsilon,$$

avec $\beta_{j,k} \in \mathbb{R}$ pour tous $j, k = 1, \ldots, d$. Plus récemment, des modèles plus complexes ont été considérés, comme des modèles linéaires généralisés (voir les références citées dans Khuri 2001) ou des modèles non-paramétriques (Facer et Müller, 2003), de façon à améliorer l'estimation de la surface $y = m(x_1, \ldots, x_d)$, en particulier lorsque la fonction m est régulière. Notons que plus le modèle est complexe et plus le nombre d'expériences à réaliser pour estimer ses paramètres correctement est important.

De façon analogue au choix du modèle de régression, le choix du plan d'expériences doit répondre à deux critères antagonistes : d'une part le nombre d'expériences à réaliser doit être le plus petit possible de façon à minimiser les coûts, d'autre part l'estimation de la surface doit être la plus précise possible. Les plans d'expériences classiques sont les plans factoriels, les plans composites centrés (CCD) et de Box-Benhken. De nombreux autres plans d'expériences existent, nous renvoyons à Georgiou, Stylianou et Aggarwal (2014) pour les avancées les plus récentes sur ce sujet et à Khuri et Mukhopadhyay (2010) pour un état de l'art et la description des plans les plus classiques.

1.5.2 Motivations pour une adaptation à un contexte fonctionnel

La méthode des surfaces de réponse s'est révélée extrêmement utile dans de nombreux domaines. Plus récemment, avec le développement de codes de calcul qui peuvent être très consommateurs en temps, cette méthode a été étendue aux expérimentations numériques (Sacks et al., 1989) et a été largement utilisée dans l'industrie, par exemple pour optimiser la conception de produits manufacturés, comme des circuits électriques (Bates et al., 1996) ou des pales de rotor (Lee et Hajela, 1996).

Cependant, aucune méthode n'existe pour optimiser une variable de sortie lorsqu'une des covariables est une fonction ou une courbe. Or les besoins sont réels tant du point de vue de l'optimisation de sorties de codes de calcul que de résultats d'expériences physiques.

Par exemple, la probabilité de défaillance de la cuve d'un réacteur nucléaire lors d'un accident de perte de réfrigérant primaire est liée à l'évolution de la température, de la pression et de l'effusivité thermique (qui mesure la capacité du matériau à échanger de la chaleur avec son environnement par contact) dans le cœur du réacteur. Ces paramètres sont contrôlés par l'injection d'eau froide dans le circuit primaire. Une meilleure compréhension de ce lien permettrait d'améliorer la procédure à suivre pour minimiser les dégâts causés par ce type d'accident.

Hormis les applications industrielles, il est possible d'envisager d'autres applications,

comme des applications médicales.

1.5.3 Contributions du chapitre 4

Dans le chapitre 4, nous proposons donc une adaptation de la méthode des surfaces de réponses pour des problèmes d'optimisation d'une variable de sortie y en fonction d'une variable fonctionnelle $x \in \mathbb{H}$.

La clef de la méthode est la définition de plans d'expériences pour variables fonctionnelles. Nous considérons une méthode souple basée sur la réduction de la dimension permettant d'exploiter les bonnes propriétés des plans d'expériences multivariés classiques. Étant donné une famille orthonormée $(\varphi_j)_{1\leq j\leq d}$ de $\mathbb{H}$ et un plan d'expériences multivarié $(\mathbf{x}_i, i = 1, \ldots, n)$ avec, pour tout i, $\mathbf{x}_i = (\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,d})' \in \mathbb{R}^d$, nous générons un plan d'expériences fonctionnel $(x_i, i = 1, \ldots, n)$ de la façon suivante

$$x_i := \sum_{j=1}^{d} \mathbf{x}_{i,j}\varphi_j.$$

L'algorithme proposé est directement inspiré de la version classique, multivariée, de la méthodologie des surfaces de réponse.

1. Génération d'un plan d'expériences fonctionnel $(x_i^{(0)}, i = 1, \ldots, n_0)$ dans une certaine région de l'espace et réalisation des expériences correspondantes (les résultats sont notés $(Y_i^{(0)}, i = 1, \ldots, n_0)$).
2. Estimation des paramètres d'un modèle d'ordre 1 : $Y = \alpha + \langle \beta, x\rangle + \varepsilon$ (avec $\alpha \in \mathbb{R}$ et $\beta \in \mathbb{H}$) à l'aide de l'échantillon $((x_i^{(0)}, Y_i^{(0)}), i = 1, \ldots, n)$ obtenu à l'étape 1, et calcul d'une direction de descente.
3. Optimisation le long de la direction de descente.
4. Réalisation de nouvelles expériences autour du point optimal et estimation des coefficients d'un modèle d'ordre deux $Y = \alpha + \langle \beta, x\rangle + \langle Hx, x\rangle + \varepsilon$ (avec $\alpha \in \mathbb{R}$, $\beta \in \mathbb{H}$ et $H : \mathbb{H} \to \mathbb{H}$ un opérateur auto-adjoint) pour affiner la localisation du minimum.

Nous testons l'algorithme sur deux exemples de données simulées. Les résultats sont encourageants et mettent en évidence l'importance d'un choix approprié de la base $(\varphi_j)_{j\geq 1}$. Deux bases *data-driven*, c'est-à-dire générées à partir d'un échantillon d'apprentissage, sont considérées : la base de l'ACP et la base PLS. La base PLS qui semble la mieux adaptée à notre contexte puisqu'elle permet de prendre en compte la corrélation entre X et Y (contrairement à la base de l'ACP qui ne dépend que de X), donne de très bons résultats à condition que l'échantillon d'apprentissage soit bien choisi. La question du choix de la dimension d est également étudiée d'un point de vue pratique : nous constatons que l'algorithme fonctionne d'autant mieux que la dimension d est grande. Cela est dû, d'une part à une meilleure exploration de l'espace $\mathbb{H}$, d'autre part, au fait que nous faisons croître le nombre d'expériences réalisées avec la dimension ce qui accroît la précision de l'estimation des modèles. La contrainte sur la dimension est donc plutôt liée au nombre d'expériences qu'il est possible de réaliser.

Nous appliquons ensuite la méthode de génération d'un plan d'expériences à des données transmises par le CEA Cadarache, issues de résultats d'un code de calcul. Ces données

sont composées de courbes de température, de pression et d'effusivité thermique observées lors d'une simulation numérique d'accident de perte de réfrigérant primaire. Nous sommes en mesure de fournir un plan d'expériences à partir de ces données. La perspective de ces travaux est de générer un échantillon afin de pouvoir estimer la direction de plus forte pente. L'objectif final est de minimiser la probabilité de défaillance de la cuve lors de l'accident.

Chapitre 2

Prédiction dans le modèle linéaire fonctionnel : point de vue de la sélection de modèle

L'objectif de ce chapitre est de proposer des résultats non-asymptotiques pour des estimateurs adaptatifs de la fonction de pente dans le modèle linéaire fonctionnel.

Nous définissons dans un premier temps un estimateur des moindres carrés sur l'espace engendré par les fonctions propres de l'opérateur de covariance Γ. Le choix d'un tel espace est motivé par ses propriétés d'optimalité (voir section 1.1.1).

Cependant, dans des contextes généraux, cette base est inconnue du statisticien et doit être estimée. La procédure classique consiste à prendre les fonctions propres de l'opérateur de covariance empirique Γ_n. L'estimation par moindres carrés sur cet espace, qui est donc ici un espace aléatoire, est connue sous le nom de régression en composantes principales.

Pour ces deux estimateurs, nous proposons une procédure de sélection de la dimension par minimisation d'un critère pénalisé. Nous montrons que les deux estimateurs sélectionnés vérifient une inégalité-oracle pour l'erreur de prédiction.

Les propriétés numériques des deux estimateurs sont étudiées.

Les résultats de ce chapitre ont fait l'objet de deux travaux soumis :

- Brunel, E., Mas. A. et Roche, A. (2013). Non-asymptotic Adaptive Prediction in Functional Linear Models, HAL : hal-00763924, arXiv : 1301.3017.
- Brunel, E. et Roche, A. (2012). Penalized contrast estimation in functional linear models with circular data, HAL : hal-00651399, en révision.

Sommaire

2.1 Introduction

We recall that the Functional Linear Model (FLM) assumes a linear dependence between a real-valued response Y and a functional predictor X belonging to an infinite-dimensional separable Hilbert space $(\mathbb{H}, <\cdot,\cdot>, \|\cdot\|)$ given by

$$Y =< \beta, X > +\varepsilon, \tag{2.1}$$

where ε stands for a centered noise term with unknown variance σ^2 and is independent of X and $\beta \in \mathbb{H}$ is an unknown function to be estimated. We suppose the random variable X to be centred as well and that $E\left[\|X\|^2\right] < +\infty$.

2.1.1 Motivation

Functional linear models have proven their ability to make accurate predictions in many practical situations (see e.g. Cardot, Crambes, and Sarda 2007, for ozone peak prediction ; Cho et al. 2013, for prediction of electricity consumption ; Bayle, Monestiez, and Nerini 2014, for application to environmental data).

All the estimators of the slope function β existing in the literature rely on the choice of at least one tuning parameter (the smoothing parameter appearing in the penalized criterion or the dimension of approximation space) which influences significantly the quality of estimation. Optimal choice of such parameters depends generally on both unknown regularities of the slope function β and the predictor X (see e.g. Cai and Hall, 2006; Crambes, Kneip, and Sarda, 2009; Cardot and Johannes, 2010 and also Section 1.3 p.13) and the parameters are usually chosen in practice by cross-validation.

Until the recent work of Comte and Johannes (2010), nonasymptotic results providing adaptive data-driven estimators were missing. Comte and Johannes (2010) propose to select the dimension of the orthogonal series estimator introduced first by Cardot and Johannes (2010) by minimization of a penalized contrast criterion under the assumption that the data X is *circular* in $\mathbb{L}^2([0,1])$ that is to say $X(0) = X(1)$ and X is second-order stationary. Their results are generalized in Comte and Johannes (2012) to non-circular data. The

dimension is selected by means of a stochastic penalized contrast criterion emulating Lepski's method (Goldenshluger and Lepski, 2011). Both resulting estimators are completely data-driven and achieve optimal minimax rates for general weighted $\mathbb{L}^2$-risks. However, the quality of an estimator $\widehat{\beta}$ of the slope β can also be evaluated in terms of prediction. Some of the literature focus on prediction error (Crambes, Kneip, and Sarda, 2009; Cardot and Johannes, 2010) which is defined by

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{Y}_{n+1} - \mathbb{E}[Y_{n+1}|X_{n+1}]\right)^2 |(X_1,Y_1),...,(X_n,Y_n)\right] &= \|\Gamma^{1/2}(\widehat{\beta}-\beta)\|^2 =: \|\widehat{\beta}-\beta\|_\Gamma^2, \\ &= \sum_{j\geq 1} \lambda_j \langle \widehat{\beta}-\beta, \psi_j\rangle^2. \qquad (2.2)\end{aligned}$$

This quantity (2.2) defines a norm on $\mathbb{H}$ denoted by $\|\cdot\|_\Gamma^2$ as soon as $\lambda_j > 0$ for all $j \geq 1$. Since this assumption on the positivity of the eigenvalues $(\lambda_j)_{j\geq 1}$ is necessary for the model to be identifiable (see Section 1.3.2, p.13), we suppose in the sequel that it is verified.

The approach of Comte and Johannes (2010, 2012) do not recover prediction error since their dimension selection criteria depends on the weights defining the risk. For prediction error these weights are the eigenvalues of the covariance operator Γ, which are unknown in practice.

Another approach is proposed by Cai and Yuan (2012) who develop a data-driven choice of the tuning parameter of the roughness regularization method (Ramsay and Silverman, 2005) by estimating directly the optimal order of the estimation parameter. The function to estimate β is supposed to lie in a reproducing kernel Hilbert space (RKHS). The mean excess risk of their estimator is shown to attain the optimal rate of convergence which depends both on the covariance operator associated to X and on the reproducing kernel. However, the limitation of their method is to require that the kernel of the associated RKHS is known which amounts to suppose that the regularity of the function to estimate β is known.

The motivation of this work is to propose an entirely data-driven procedure to select the adequate dimension D_m for least-squares estimators defined on both approximation spaces S_{D_m} (when it is known) and $\widehat{S}_{D_m}$ and to give non-asymptotic results in terms of prediction error.

2.1.2 Organisation of the chapter

The estimation procedures are presented in Section 2.2. The resulting estimators are proved to satisfy an oracle-type inequality and to attain the optimal minimax rate of convergence for the risk associated to the prediction error for slope functions belonging to Sobolev classes in Section 2.3. In Section 2.4, a simulation study is presented including a comparison with cross-validation. The proofs are detailed in Section 2.5 and in Appendix A.

2.2 Penalized contrast estimation of the slope function

2.2.1 Least-squares estimation

In this section we fix a dimension D_m and define estimators in both spaces $S_m = \text{span}\{\psi_1, \ldots, \psi_{D_m}\}$ (when the eigenfunctions $(\psi_j)_{j\geq 1}$ of Γ are known) and

$\widehat{S}_m = \text{span}\{\widehat{\psi}_1, \ldots, \widehat{\psi}_{D_m}\}$.

Case where the eigenfunctions of Γ are known

We define — in case that this definition makes sense — the least squares estimator $\widehat{\beta}_m^{(KB)}$ of β in S_m by:

$$\widehat{\beta}_m^{(KB)} := \arg\min_{f \in S_m} \gamma_n(f), \tag{2.3}$$

where

$$\gamma_n(f) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - < f, X_i >)^2. \tag{2.4}$$

The function $f = \sum_{j=1}^{D_m} \alpha_j \psi_j$ minimizes the contrast γ_n on S_m if and only if the vector $(\alpha_1, ..., \alpha_{D_m})' \in \mathbf{R}^{D_m}$ minimizes the convex function

$$F(t_1, ..., t_{D_m}) := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{D_m} t_j < \psi_j, X_i >\right)^2$$

on $\mathbf{R}^{D_m}$. Let us define the matrix

$$\widehat{\Phi}_m := (\langle \Gamma_n \psi_j, \psi_k \rangle)_{1 \le j,k \le D_m} \tag{2.5}$$

and the vector

$$g_m = \left(\frac{1}{n}\sum_{i=1}^{n} Y_i < \psi_j, X_i >\right)'_{1 \le j \le D_m},$$

we have:

$$\nabla F(t) = -2g_m + 2\widehat{\Phi}_m t,$$

with $t = (t_1, \cdots, t_{D_m})' \in \mathbf{R}^{D_m}$.

Therefore, we have existence and uniqueness of the least squares estimator $\widehat{\beta}_m^{(KB)}$ on S_m if and only if the matrix $\widehat{\Phi}_m$ is invertible. In that case, the estimator is given by:

$$\widehat{\beta}_m^{(KB)} = \sum_{j=1}^{D_m} \alpha_j \psi_j,$$

with $\alpha = \widehat{\Phi}_m^{-1} g_m$.

Case where the eigenfunctions of Γ are unknown – Functional Principal Component Regression (FPCR)

Define $\widehat{g} := (1/n)\sum_{i=1}^{n} Y_i X_i$ the cross-covariance between Y and X and

$$\widehat{\beta}_m^{(FPCR)} := \sum_{j=1}^{D_m} \frac{< \widehat{g}, \widehat{\psi}_j >}{\widehat{\lambda}_j} \widehat{\psi}_j. \tag{2.6}$$

We can see easily that (2.6) is the unique minimizer of the least-squares contrast γ_n defined by Equation (2.4) if $\widehat{\lambda}_m > 0$.

In the sequel, when a property applies to both $\widehat{\beta}_m^{(KB)}$ and $\widehat{\beta}_m^{(FPCR)}$ we denote simply these estimators by $\widehat{\beta}_m$. In the next section, we define a model selection criterion to select an estimator in the family $\left\{\widehat{\beta}_m,\ m \in \widehat{\mathcal{M}}_n\right\}$ where $\widehat{\mathcal{M}}_n$ is a data-driven model collection.

2.2.2 Model selection criterion

We give here some additional details on the link between m and the dimension of S_m, denoted by D_m. In all the chapter, the sequence $(D_m)_{m\geq 1}$ is an increasing sequence of positive integers. For the estimator $\widehat{\beta}^{(FPCR)}$, we set $D_m = m$. However, for the estimator $\widehat{\beta}^{(KB)}$ it could be useful to restrict the admissible dimensions: for instance when X is second-order stationary and $X(0) = X(1)$ almost surely (circular data), the space S_m is spanned by the Fourier basis and it seems more appropriate to set $D_m = 2m + 1$. For this reason, and also to keep unified notations for both estimators $\widehat{\beta}^{(FPCR)}$ and $\widehat{\beta}^{(KB)}$, we choose the general notation D_m.

Since the theoretical results are true if the maximal dimension we can consider satisfies some constraints depending on the sequence $(\lambda_j)_{j\geq 1}$ (which is unknown), we define empirical maximal dimensions.

Let $\theta > 4$ and $\delta > 0$ and $\mathfrak{s}_n := \frac{2}{n^2}(1 - 1/\sqrt{\ln(n)})$,

$$\widehat{N}_n^{(FPCR)} := \max\left\{N \in \mathbb{N}^*, D_N \leq \min\{20\sqrt{n/\ln^3(n)}, n/\theta(1+2\delta)\} \text{ and } \widehat{\lambda}_{D_N} \geq \mathfrak{s}_n\right\},$$

where we recall that $(\widehat{\lambda}_j)_{j\geq 1}$ are the eigenvalues of the empirical covariance operator Γ_n.

We also define,

$$N_n := \max\left\{N \in \mathbb{N}^*, D_N \leq \min\{20\sqrt{n/\ln^3(n)}, n/\theta(1+2\delta)\} \text{ and } \lambda_{D_N} \geq n^{-2}\right\}.$$

We set $\widehat{\mathcal{M}}_n^{(FPCR)} := \left\{1, \ldots, \widehat{N_n}^{(FPCR)}\right\}$ and $\mathcal{M}_n := \{1, \ldots, N_n\}$.

The dimension selection criterion is very similar for both estimators: we define

$$\widehat{m}^{(KB)} := \arg\min_{m\in\mathcal{M}_n}(\mathrm{crit}(m))$$
$$\widehat{m}^{(FPCR)} := \arg\min_{m\in\widehat{\mathcal{M}_n}}(\mathrm{crit}(m))$$

where

$$\mathrm{crit}(m) := \gamma_n(\widehat{\beta}_m) + \widehat{\mathrm{pen}}(m) = \gamma_n(\widehat{\beta}_m)\left(1 + \theta(1+\delta)\frac{D_m}{n}\right), \tag{2.7}$$

$$\widehat{\mathrm{pen}}(m) := \theta(1+\delta)\widehat{\sigma}_m^2\frac{D_m}{n},$$

and $\widehat{\sigma}_m^2$ is a plug-in estimator of the noise variance σ^2 defined by

$$\widehat{\sigma}_m^2 := \frac{1}{n}\sum_{i=1}^{n}(Y_i - < \widehat{\beta}_m, X_i >)^2 = \gamma_n(\widehat{\beta}_m).$$

We replace $\widehat{\beta}_m$ by either $\widehat{\beta}_m^{(KB)}$ of $\widehat{\beta}_m^{(FPCR)}$ where required.

Then, we propose the following estimators of the function β

$$\begin{aligned}\widetilde{\beta}^{(KB)} &:= \widehat{\beta}^{(KB)}_{\widehat{m}^{(KB)}}\mathbf{1}_{\overline{G}}\\ \widetilde{\beta}^{(FPCR)} &:= \widehat{\beta}^{(FPCR)}_{\widehat{m}^{(FPCR)}}\end{aligned}$$

where

$$\overline{G} = \bigcap_{m\in\mathcal{M}_n}\left\{2\widetilde{\lambda}_m \geq \mathfrak{s}_n\right\}$$

with $\widetilde{\lambda}_m$ the smallest eigenvalue of $\widehat{\Phi}_m$. We simply write $\widetilde{\beta}$ when a property applies to both estimators.

Remark that Baraud, Giraud, and Huet (2009) also propose a multiplicative criterion for model selection purpose in the Gaussian regression framework with unknown noise variance. Translated to our context, with our notations, this consists in taking $\widehat{\sigma}_m^2 = \frac{n}{n-D_m}\gamma_n(\widehat{\beta}_m)$.

For the FPCR estimator, a major difference appears here with classical model selection device. Indeed, they are always carried out with fixed and known model collections. Here we handle random bases and this is the source of additional problems related to the convergence of (possibly random) projectors associated to these finite-dimensional spaces. Other difficulties come from the non-linear dependency between the coefficients of our estimator in the basis $(\widehat{\psi}_j, ..., \widehat{\psi}_m)$ and the basis itself. Solving these problems requires specific tools. Here we have used tools from perturbation theory. The presentation of this theory as well as the proofs relying on it are given in Appendix A.

2.3 Upper-bound for the prediction error

2.3.1 Assumptions

Recall that $(\lambda_j, \psi_j)_{j\geq 1}$ denote the eigenelements of the covariance operator Γ. We first suppose that all the eigenvalues $(\lambda_j)_{j\geq 1}$ are distinct, consequently $\lambda_1 > \lambda_2 > \ldots$. This assumption may be relaxed but at the price of technical difficulties.

We can control the risk under four additional assumptions:

H1 There exists $p > 4$ such that $\tau_p := \mathbb{E}[|\varepsilon|^p] < +\infty$.

H2 There exists $b > 0$ such that, for all $\ell \in \mathbb{N}^*$,

$$\sup_{j\in\mathbb{N}}\mathbb{E}\left[\frac{< X, \psi_j >^{2\ell}}{\lambda_j^\ell}\right] \leq \ell! b^{\ell-1}.$$

In order to deal with random approximation spaces $\left\{\widehat{S}_m, m \in \widehat{\mathcal{M}}_n\right\}$, we will need in addition the two following assumptions

H3 For all $j \neq k$, $< X, \psi_j >$ is independent of $< X, \psi_k >$.

H4 There exists a constant $\gamma > 0$ such that the sequence $\left(j\lambda_j \max\{\ln^{1+\gamma}(j), 1\}\right)_{j\geq 1}$ is decreasing.

Assumption **H1** is standard in regression, it is verified for most of the standard distributions (uniform, normal,...). Assumption **H2** is necessary to apply exponential inequalities, it is verified for Gaussian processes. Assumption **H3** is also classical and we know from the Karhunen-Loeve decomposition of X that it is true for Gaussian processes too (see Ash and Gardner 1975, Section 1.4). Moreover, note that for every general random variables $X \in \mathbb{H}$, the random variables $< X, \psi_j >$ and $< X, \psi_k >$ are uncorrelated since if $j \neq k$, $\mathbb{E}[< \psi_j, X_i >< \psi_k, X_i >] =< \Gamma\psi_j, \psi_k >= 0$. The assumption **H4** on the sequence $\left(j\lambda_j \max\{\ln^{1+\gamma}(j), 1\}\right)_{j\geq 1}$ allows to avoid more restrictive hypotheses about spacing control between eigenvalues as usually made in the literature (see Cai and Hall, 2006, Hall and Hosseini-Nasab, 2006, Hall and Horowitz, 2007).

2.3.2 Oracle-type inequality for the empirical risk

We define an empirical semi-norm naturally associated to our estimation problem by

$$\|f\|^2_{\Gamma_n} := \|\Gamma_n^{1/2} f\|^2 = \frac{1}{n}\sum_{i=1}^{n} < f, X_i >^2, \text{ for all } f \in \mathbb{H}.$$

In a first step, in Proposition 1, we prove that our estimators verify an oracle type inequality for the risk associated to this semi-norm whatever the regularity of the slope function β and the decreasing rate of the covariance operator eigenvalues are.

For all m, we denote by Π_m the orthogonal projection operator on S_m and by $\widehat{\Pi}_m$ the orthogonal projection operator on $\widehat{S}_m$.

Proposition 1. *Suppose that Assumption **H1** is fulfilled. We have, if $n \geq 2$,*

$$\mathbb{E}[\|\widetilde{\beta}^{(KB)} - \beta\|^2_{\Gamma_n}] \leq C\left(\inf_{m\in\mathcal{M}_n}\left\{\mathbb{E}\left[\|\beta - \Pi_m\beta\|^2_{\Gamma_n}\right] + \text{pen}(m)\right\} + \frac{\tau_p^{2/p} + \|\beta\|^2}{n}\right), \quad (2.8)$$

and

$$\mathbb{E}[\|\widetilde{\beta}^{(FPCR)} - \beta\|^2_{\Gamma_n}] \leq C'\left(\inf_{m\in\mathcal{M}_n}\left\{\mathbb{E}[\|\beta - \widehat{\Pi}_m\beta\|^2_{\Gamma_n}] + \text{pen}(m)\right\} + \frac{\tau_p^{2/p} + \|\beta\|^2}{n}\right), \quad (2.9)$$

where

$$\text{pen}(m) := \theta(1+\delta)\sigma^2\frac{D_m}{n},$$

is the theoretical version of $\widehat{\text{pen}}(m)$, $\mathcal{M}_n := \{1, \ldots, N_n\}$ the theoretical version of $\widehat{\mathcal{M}}_n$, and $C, C' > 0$ depend only on θ, p and δ.

2.3.3 Oracle-type inequality for the prediction error

In this section, we derive an oracle-inequality for the risk associated to the prediction error. For $\widetilde{\beta}^{(FPCR)}$, this oracle-inequality is obtained at the price of additional assumptions. One of them is to suppose that β is in an ellipsoid of $\mathbb{H}$, denoted $\mathcal{W}_r^R$, and depending on

two real numbers $r, R > 0$,

$$\mathcal{W}_r^R := \left\{ f \in \mathbb{H},\ \sum_{j\geq 1} j^r < f, \psi_j >^2 \leq R^2 \right\}.$$

In the case where $\mathbb{H} = \mathbb{L}^2([0,1])$ and $(\psi_j)_{j\geq 1}$ is the Fourier basis, a function β belongs to $\mathcal{W}_r^R$ if it is sufficiently "smooth". More precisely, for any even integer $r = 2k \in \mathbb{N}^*$, $\beta \in \mathcal{W}_r^R$ if and only if:

- the $(k-1)$-th derivative of β, $\beta^{(k-1)}$ is absolutely continuous;
- there exists $L > 0$ such that $\|\beta\| \leq L$;
- for all $j = 1, \ldots, k-1$, $\beta^{(j)}(0) = \beta^{(j)}(1)$.

This result is due to Tsybakov (2009, Lemma A.3).

We also need to precise the decreasing rate of the sequence $(\lambda_j)_{j\geq 1}$. Usually in functional linear regression, this rate is supposed to be polynomial (see for instance Cai and Hall 2006; Crambes, Kneip, and Sarda 2009; Cai and Yuan 2012) but more regular processes may be considered. That is the reason why, following Cardot and Johannes (2010) or Comte and Johannes (2010), we also consider exponential rates.

(P) Polynomial decrease. There exists two constants $a > 1$ and $c_P \geq 1$ such that, for all $j \geq 1$

$$c_P^{-1} j^{-a} \leq \lambda_j \leq c_P j^{-a}.$$

(E) Exponential decrease. There exists two constants $a > 0$ and $c_E \geq 1$ such that for all $j \geq 1$

$$c_E^{-1} \exp(-j^a) \leq \lambda_j \leq c_E \exp(-j^a).$$

Theorem 1. *Suppose that both assumptions **H1** and **H2** hold, if* $\mathbb{E}[\|X\|^4] < +\infty$,

$$\mathbb{E}[\|\widetilde{\beta}^{(KB)} - \beta\|_\Gamma^2] \leq C_1 \left(\min_{m\in\mathcal{M}_n} \left(\|\beta - \Pi_m \beta\|_\Gamma^2 + \mathrm{pen}(m) \right) \right) + \frac{C_2(\tau_p^{2/p} + \|\beta\|^2)}{n}, \qquad (2.10)$$

where the constants $C_1 > 0$ and $C_2 > 0$ do not depend on β or n.

*Suppose that, in addition, assumptions **H3** and **H4** hold and that the decreasing rate of $(\lambda_j)_{j\geq 1}$ is given by **(P)** or **(E)**. Suppose moreover that $\beta \in \mathcal{W}_r^R$ and that, in the polynomial case **(P)**, $a + r/2 > 2$. We have, if $n \geq 6$:*

$$\mathbb{E}[\|\widetilde{\beta}^{(FPCR)} - \beta\|_\Gamma^2] \leq C_1' \left(\min_{m\in\mathcal{M}_n} \left(\mathbb{E}[\|\beta - \widehat{\Pi}_m \beta\|_\Gamma^2] + \mathrm{pen}(m) \right) \right) + \frac{C_2'}{n}, \qquad (2.11)$$

where the constants $C_1' > 0$ and $C_2' > 0$ do not depend on β or n.

Inequality (2.10) is obtained from Inequality (2.8) by a control of the quantity $\inf_{f\in S_m} \frac{\|f\|_{\Gamma_n}^2}{\|f\|_\Gamma^2}$ which is equal to the minimal eigenvalue of a random matrix. The proof of Inequality (2.11) follows the same idea. However, additional difficulties come from the randomness of the space $\widehat{S}_m$.

- The quantity $\inf_{f\in\widehat{S}_m} \frac{\|f\|_{\Gamma_n}^2}{\|f\|_\Gamma^2}$ is harder to control than $\inf_{f\in S_m} \frac{\|f\|_{\Gamma_n}^2}{\|f\|_\Gamma^2}$.

– For the non-random projector Π_m, we have

$$\mathbb{E}\left[\|\beta - \Pi_m\beta\|^2_{\Gamma_n}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\langle\beta - \Pi_m\beta, X_i\rangle^2\right] = \|\beta - \Pi_m\beta\|^2_{\Gamma}.$$

This property is no longer true for $\widehat{\Pi}_m$ since, in general, $\mathbb{E}\left[\langle\beta - \widehat{\Pi}_m\beta, X_1\rangle^2\right] \neq \mathbb{E}\left[\|\beta - \widehat{\Pi}_m\beta\|^2_{\Gamma}\right]$. The upper-bound on $\mathbb{E}\left[\|\beta - \widehat{\Pi}_m\beta\|^2_{\Gamma_n}\right]$ is obtained in Lemma 19 in Appendix A p.170.

Remark 1: The condition $a + r/2 > 2$ is verified as soon as $a \geq 2$ without condition on the regularity parameter r of the slope β. Note that if X is a Brownian motion, $\lambda_j = \pi^{-2}(j - 0.5)^{-2}$ (Ash and Gardner, 1975). Hence, Assumption **(P)** is verified with $a = 2$ and Assumption **H4** is verified with e.g. $\gamma = 1$. Hence, the assumptions of Theorem 1 are fulfilled regardless the value of r.

2.3.4 Convergence rates

As a consequence of the oracle-inequality given in Theorem 1, we derive uniform bounds on the risk of our estimators on the ellipsoids $\mathcal{W}_r^R$.

Theorem 2. *Assume that the assumptions of Theorem 1 are fulfilled. For all $r > 0$ and $R > 0$:*

Polynomial case. *If **(P)** holds:*

$$\sup_{\beta\in\mathcal{W}_r^R}\mathbb{E}[\|\widetilde{\beta} - \beta\|^2_{\Gamma}] \leq C_P n^{-(a+r)/(a+r+1)}. \tag{2.12}$$

Exponential case. *If **(E)** holds then:*

$$\sup_{\beta\in\mathcal{W}_r^R}\mathbb{E}[\|\widetilde{\beta} - \beta\|^2_{\Gamma}] \leq C_E n^{-1}(\ln n)^{1/a}, \tag{2.13}$$

with C_P and C_E independent of n.

Remark 2: In the case where the noise ε is Gaussian, the bounds (2.12) and (2.13) coincide, up to the multiplicative constants, with the minimal bounds given by Cardot and Johannes (2010). Similar rates are obtained by Cai and Yuan (2012) for the excess risk which is very close to the prediction error.

2.4 Numerical results

2.4.1 Sample simulation

In order to provide a method of simulation of the curve X, we start from the Karhunen-Loève decomposition of X,

$$X = \sum_{j\geq 1} < X, \psi_j > \psi_j,$$

where we recall that the convergence of the series takes place in $\|\cdot\|$.

The sequence $(< X, \psi_j >)_{j\geq 1}$ is a sequence of uncorrelated and centred random variables with variance $\mathbb{E}[< X, \psi_j >^2] =< \Gamma\psi_j, \psi_j >= \lambda_j$.
We are taking up the idea of Hall and Hosseini-Nasab (2006) to simulate X using a truncated version of its Karhunen-Loève decomposition at a rank $J \in \mathbb{N}^*$:

$$X(t) := \sum_{j=1}^{J} \sqrt{\lambda_j}\xi_j\psi_j(t). \tag{2.14}$$

This allows to explicitly choose the eigenvalues of the operator Γ and, in particular, to study the influence of the decreasing rate of $(\lambda_j)_{j\geq 1}$ on the quality of estimation. The rank J has to be chosen sufficiently large such that $\sum_{j>J}\lambda_j$ is negligible with respect to $\sum_{j=1}^{J}\lambda_j$. In the sequel, we take $J > 100$.

We simulate a Gaussian process, then here $(\xi_j)_{1\leq j\leq J}$ is a standard normal Gaussian vector.

In the following, we set $\mathbb{H} = \mathbb{L}^2([0,1])$ and, except in Section 2.4.5, we will consider the following family of basis:

$$\psi_j^{(1)} : t \in [0,1] \mapsto \sqrt{2}\sin(\pi(j-0.5)t), \text{ for all } j \geq 1.$$

This choice of basis functions corresponds to the family of eigenfunctions associated to the Brownian motion (see Ash and Gardner 1975), hence we can simulate a Brownian motion by taking $\lambda_j = (j-0.5)^{-2}\pi^{-2}$.

In order to study the effect of the estimation of the eigenfunctions on the quality of estimation, we will compare both estimators $\widetilde{\beta}^{(KB)}$ and $\widetilde{\beta}^{(FPCR)}$ and simulate circular data. We consider the Fourier basis:

$$\psi_1^{(2)} \equiv 1,\ \psi_{2j}^{(2)} : t \in [0,1] \mapsto \sqrt{2}\cos(2\pi t) \text{ and } \psi_{2j+1}^{(2)} : t \in [0,1] \mapsto \sqrt{2}\sin(2\pi t).$$

In the following, we set, for $i = 1, 2$,

$$X^{(i)}(t) := \sum_{j=1}^{J} \sqrt{\lambda_j}\xi_j\psi_j^{(i)}(t).$$

Examples of simulated curves are given in Figure 2.1.

To match with what happens for real data, the random function X is discretized and we treat the sequence $(X(t_1), ..., X(t_p))$ with $p \in \mathbb{N}^*$ (hereafter $p = 100$ and $t_j := \frac{j-1}{p-1}$ for all $j = 1, ..., p$).

Then Y_i is obtained directly from its definition and the scalar product is approximated by the trapezoidal rule

$$\begin{aligned} Y_i &= < \beta, X_i > +\varepsilon_i \\ &\approx \frac{1}{p-1}\left(\frac{\beta(t_1)X(t_1) + \beta(t_p)X(t_p)}{2} + \sum_{j=2}^{p-1}\beta(t_j)X(t_j)\right) + \varepsilon_i, \end{aligned}$$

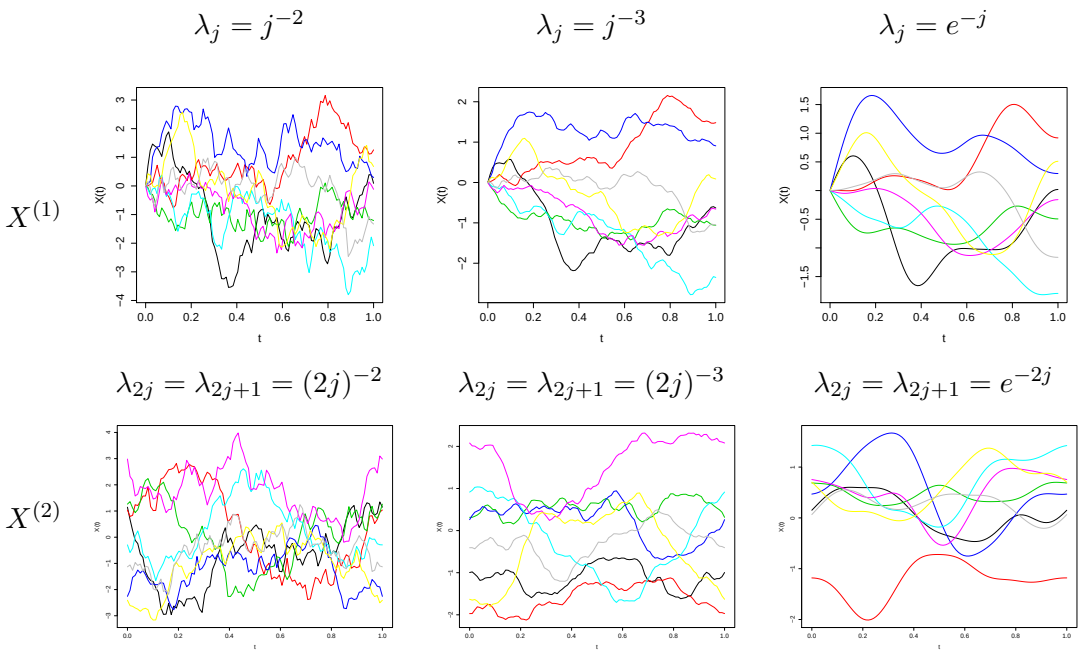


Figure 2.1: realisation of X following (2.14).

with $\varepsilon \sim \sigma\mathcal{U}([-\sqrt{3}, \sqrt{3}])$ and $\sigma^2 = 0.01$.

We mainly test two slope functions

$$\beta_1(x) = \ln(15x^2 + 10) + \cos(4\pi x) \text{ (Cardot, Ferraty, and Sarda, 1999)}$$

and

$$\beta_2(x) = e^{-(x-0.3)^2/0.05} \cos(4\pi x).$$

2.4.2 Estimation of the PCA basis

In practice, the calculation of $\widetilde{\beta}^{(KB)}$ does not cause difficulties, it is made following the procedure explained in Section 2.2.1 where we fix $D_m = 2m + 1$. For $\widetilde{\beta}^{(FPCR)}$, we take $D_m = m$ and the procedure, detailed in Section 2.2.1, is relatively similar but the calculation of the PCA basis is a practical problem. Recall that the aim is to estimate the eigenfunctions $(\widehat{\psi}_j)_{j\geq 1}$ and the eigenvalues $(\widehat{\lambda}_j)_{j\geq 1}$ of the empirical covariance operator

$$\Gamma_n : f \mapsto \frac{1}{n}\sum_{i=1}^{n} \langle f, X_i \rangle X_i.$$

While the empirical covariance operator can be explicitly calculated, the calculation of its eigenfunctions is a real difficulty. The usual method (see Ramsay and Silverman 2005, Section 8.4.2) consists in reducing the problem to a finite-dimensional linear algebra problem by calculating the eigenfunctions and eigenvalues of the restriction of Γ_n to a finite-dimensional space.

For that purpose, we choose a finite orthonormal family of function of $\mathbb{H}$, here an his-

togram basis with bin width $1/(p-1)$:

$$e_j := \sqrt{p-1}\ \mathbf{1}_{[t_j, t_{j+1}[},\ \text{for } j = 1, \dots, p-1,$$

where $(t_j)_{j=1,\dots,p-1}$ is the discretisation grid used for X. With this choice of basis, this approach is very similar to the discretisation approach described in Ramsay and Silverman (2005, Section 8.4.1).

The matrix

$$\mathbf{\Gamma} := (< \Gamma_n e_j, e_k >)_{1 \le j,k \le p-1}$$

coincides with the restriction of Γ_n to the space $\text{span}\{e_1, \dots, e_{p-1}\}$. Let $(\widehat{\lambda}_j)_{1 \le j \le p-1}$ (resp. $(V_1, ..., V_{p-1})$) the eigenvalues (resp. eigenvectors) of the matrix $\mathbf{\Gamma}$, we estimate the value of $\widehat{\psi}_j$ for $j = 1, \dots, p-1$ at a point $t \in [0,1]$ in the following way :

$$\widehat{\psi}_j(t) := \sum_{k=1}^{p-1} (V_j)_k e_k(t).$$

Figure 2.2 presents the result of the estimation of the eigenfunctions, we remark that, for all $j = 1, \dots, 8$, we estimate accurately either $\psi_j^{(1)}$ or $-\psi_j^{(1)}$ (since the eigenfunctions are defined up to the multiplication by -1).

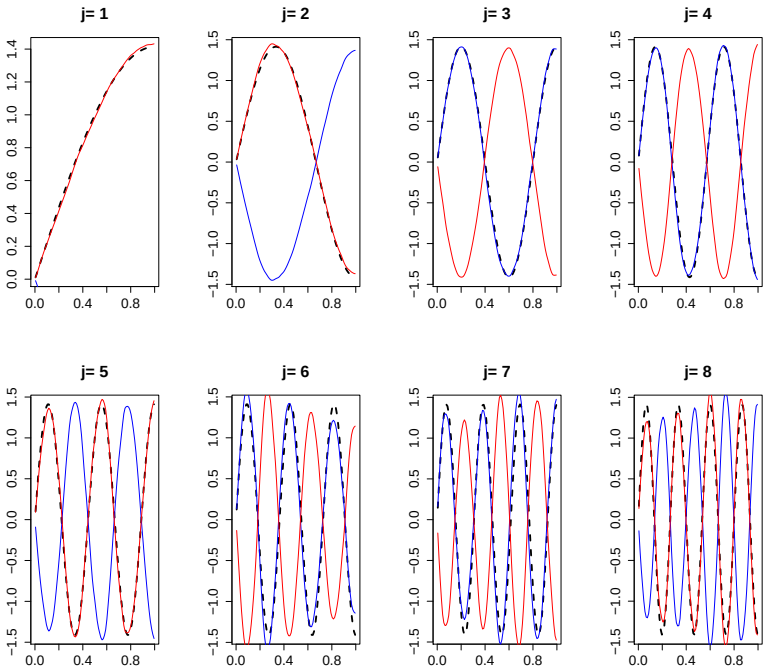


Figure 2.2: Estimation of $\psi_j^{(1)}$ for $j = 1, \dots, 8$. Black dotted curve: plot of $\psi_j^{(1)}$, blue curve: plot of $\widehat{\psi}_j^{(1)}$, red curve: plot of $-\widehat{\psi}_j^{(1)}$, $n = 1000$, $\lambda_j = j^{-3}$.

2.4.3 Calibration of the constant κ appearing in the penalty

Recall the definition of $\widetilde{\beta}$:

$$\widetilde{\beta}^{(KB)} = \widehat{\beta}^{(KB)}_{\widehat{m}^{(KB)}} \text{ and } \widetilde{\beta}^{(FPCR)} = \widehat{\beta}^{(FPCR)}_{\widehat{m}^{(FPCR)}}$$

where $\widehat{m}^{(KB)}$ minimizes the criterion

$$\text{crit}^{(KB)}(m) = \gamma_n(\widehat{\beta}^{(KB)}_m) + \kappa \widehat{\sigma}^2_m \frac{D_m}{n}$$

and $\widehat{m}^{(FPCR)}$ minimizes

$$\text{crit}^{(FPCR)}(m) = \gamma_n(\widehat{\beta}^{(FPCR)}_m) + \kappa \widehat{\sigma}^2_m \frac{D_m}{n}.$$

It is important to fix accurately the value of κ in both criteria since it influences the complexity of selected models. We study several methods of calibration for κ: one is a method based on simulations and Monte-Carlo study and the others are based on slope heuristics.

Calibration by simulations

We start from the principle that the value of κ should not depend on the law of X, or the law of the noise ε, or β or n. The idea is then to plot $\mathbb{E}[\|\widetilde{\beta}-\beta\|^2_\Gamma$ as a function of κ, for different values of n ($n = 200$, $n = 500$, $n = 1000$), $(\lambda_j)_{j\geq 1}$ and β (β_1 and $\beta_3(t) := (t-0.2)(t-0.5)$). For each set of parameters and each value of κ, the risk $\mathbb{E}[\|\widetilde{\beta}-\beta\|^2_\Gamma]$ is approximated by the mean of $N = 500$ Monte-Carlo replications of the random variable $\|\widetilde{\beta}-\beta\|^2_\Gamma$.

From the results presented in Figure 2.3 we deduce that the optimal constant κ should lie between 2 and 4.

The final goal is to select a dimension which is as close as possible to the oracle dimension. We compare numerically the dimension selected by the criterion and the optimal dimension m^* defined by

$$m^* := \arg\min_{m\in\mathcal{M}_n}\{\|\beta - \widehat{\beta}_m\|^2_\Gamma\}.$$

This definition differs from the usual definition of oracle given in the introduction (Section 1.2.1, p.9) but is more adapted to numerical contexts since it does not involve an expectation. We call m^* the *pseudo-oracle* (or sometimes *oracle* for brevity of notation).

We decide to study the quantity $\widehat{m} - m^*$ which is called hereafter *distance to the oracle*. The closer the distance to the oracle is to 0, the better it is. A negative distance suggests that we over-penalize and hence that κ is too large and a positive distance that κ is too small.

Figure 2.4 indicates that for small values of κ ($\kappa \leq 2.5$) the criterion is unstable. This instability is no longer perceptible when $\kappa \geq 3.5$, then in the sequel we fix $\kappa = 4$. Similar considerations allow us to fix $\kappa = 4$ for $\widetilde{\beta} = \widetilde{\beta}^{(KB)}$.

Slope heuristics

Alternatives to calibration by simulation, called slope heuristics, have been introduced by Birgé and Massart (2007). It is based on two principles, supported by theoretical arguments.

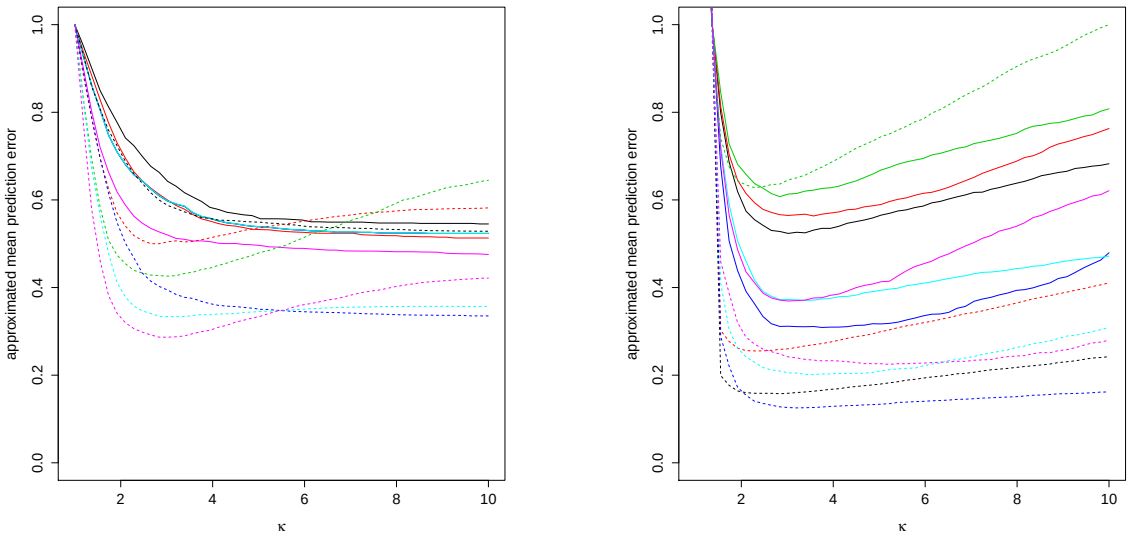


Figure 2.3: Calibration of κ. Each figure represents the approximation of $\mathbb{E}[\|\widetilde{\beta} - \beta\|_\Gamma^2]$ depending on κ for different combination of n, $(\lambda_j)_{j\geq 1}$ and β. Left-hand side: $\widetilde{\beta} = \widetilde{\beta}^{(KB)}$, right-hand side: $\widetilde{\beta} = \widetilde{\beta}^{(FPCR)}$. Legend: for $\lambda_j^{(P)}$: black curves: $n = 200$, red curves: $n = 500$, green curves: $n = 1000$; for $\lambda_j^{(E)}$: dark blue curves: $n = 200$, light blue curves: $n = 500$, pink curves: $n = 1000$; $\lambda_{2j}^{(P)} = \lambda_{2j+1}^{(P)} = (2j)^{-2}$ on the right-hand side, $\lambda_j^{(P)} = j^{-2}$ on the left-hand side, $\lambda_{2j}^{(E)} = \lambda_{2j+1}^{(E)} = e^{-2j}$ on the right-hand side, $\lambda_j^{(E)} = e^{-2j}$. For β_1, the curves are solid, for β_3, the curves are dotted. The values of each curve are first divided by their maximum, this does not change the behaviour of the curve but make the figures more visible.

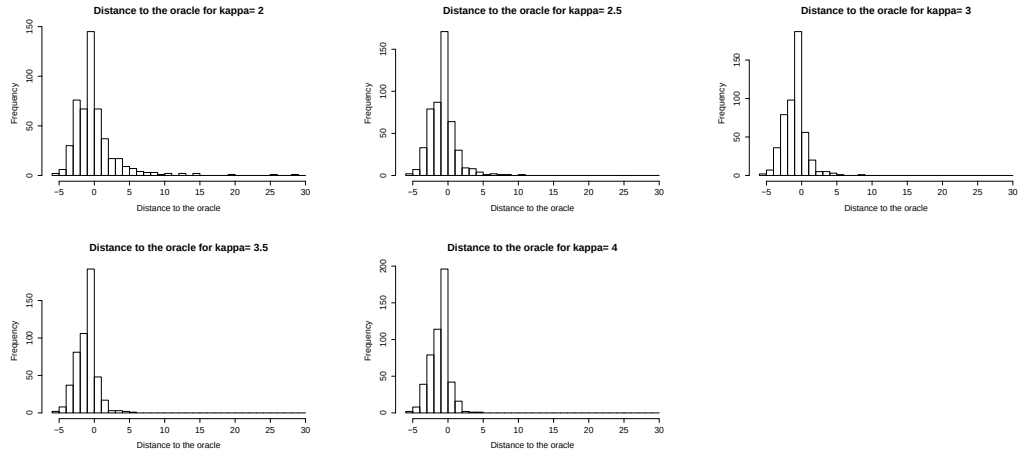


Figure 2.4: Distribution of the distance to the oracle over 500 estimators of β_2 for different values of κ, $n = 1000$, $X = X^{(1)}$, $\widetilde{\beta} = \widetilde{\beta}^{(FPCR)}$.

The first principle is the existence of a minimal penalty that is a penalty defined such that

the complexity of estimators defined with lighter penalty explodes while estimators defined with a higher penalty have a reasonable complexity. The second principle is that the ideal penalty is equal to twice the minimal penalty. The method has then been developed and extended to various statistical models both in a theoretical and practical point of view (see Baudry, Maugis, and Michel 2012 for a recent overview). The two main methods – slope estimation method and the dimension jump method – are implemented in the package R *Capushe* (Brault et al., 2012).

To understand the dimension jump method, let us detail the arguments leading to the first principle that is the existence of a minimal penalty. We give only here the arguments necessary to understand the basis of the slope heuristic method and we refer to Birgé and Massart (2007), Arlot and Massart (2009), Verzelen (2010), and Lerasle (2012) for details and theoretical validation in different contexts. Hence the following is purely heuristic and we de not pretend to make a proof. For simplicity, we suppose in this section that the noise variance σ^2 is known, in that case the penalty is written

$$\text{pen}(m) = \kappa\sigma^2 \frac{D_m}{n}. \tag{2.15}$$

Let $\text{pen}_{\min}(m) := \gamma_n(\widehat{\Pi}_m\beta) - \gamma_n\left(\widehat{\beta}_m^{(FPCR)}\right)$, we see that $\text{pen}_{\min}$ is a minimal penalty. We assume here that $\text{pen}_{\min}$ can be written in the form $\kappa_{\min}\sigma^2\frac{D_m}{n}$ (in practice it is only close to a quantity which can be written on this form). Indeed, we can describe what appears for the FPCR estimator with the following heuristics.

- Suppose that $\text{pen}(m) < \text{pen}_{\min}(m)$, then $\gamma_n\left(\widehat{\beta}_m^{(FPCR)}\right) + \text{pen}(m) < \gamma_n(\widehat{\Pi}_m\beta)$. If n is large, we have $\gamma_n(\widehat{\Pi}_m\beta) \approx \mathbb{E}\left[\gamma_n(\widehat{\Pi}_m\beta)\right] \approx \sigma^2 + \|\beta - \Pi_m\beta\|_\Gamma^2$ since this last quantity is decreasing we see that the criterion to select m will be decreasing too and will select the most complex model.
- Now suppose that $\text{pen}(m) > \text{pen}_{\min}(m)$, for instance suppose that there exists $\alpha > 1$ such that $\text{pen}(m) = \alpha\text{pen}_{\min}(m)$, the criterion to select m is then written

  $$\gamma_n\left(\widehat{\beta}_m^{(FPCR)}\right) + \text{pen}(m) = \gamma_n\left(\widehat{\beta}_m^{(FPCR)}\right) + \alpha\gamma_n\left(\widehat{\Pi}_m\beta - \widehat{\beta}_m^{(FPCR)}\right),$$

  the first term is a bias term which stabilizes when m grows while the second is a variance term which explodes when m grows. The criterion will not select too complex models.

The same reasoning is true for $\widetilde{\beta}^{(KB)}$ replacing $\widehat{\Pi}_m$ by Π_m.

This argument tells us that a complexity jump is expected around the minimal penalty. Then, if we detect the value of κ for which this jump occurs, we can detect the value of the minimal penalty, and the optimal one follows from the second principle. Figure 2.5 illustrates the method.

For the slope estimation method, recall that the second principle states that the optimal penalty is $\text{pen}_{\text{opt}} := 2\text{pen}_{\min}(m)$, we have:

$$\text{pen}_{\text{opt}}(m) = 2(\gamma_n(\widehat{\Pi}_m\beta) - \gamma_n(\beta)) + 2(\gamma_n(\beta) - \gamma_n(\widehat{\beta}_m)),$$

the first term, which is a bias term, stabilizes when m is sufficiently large, for this values

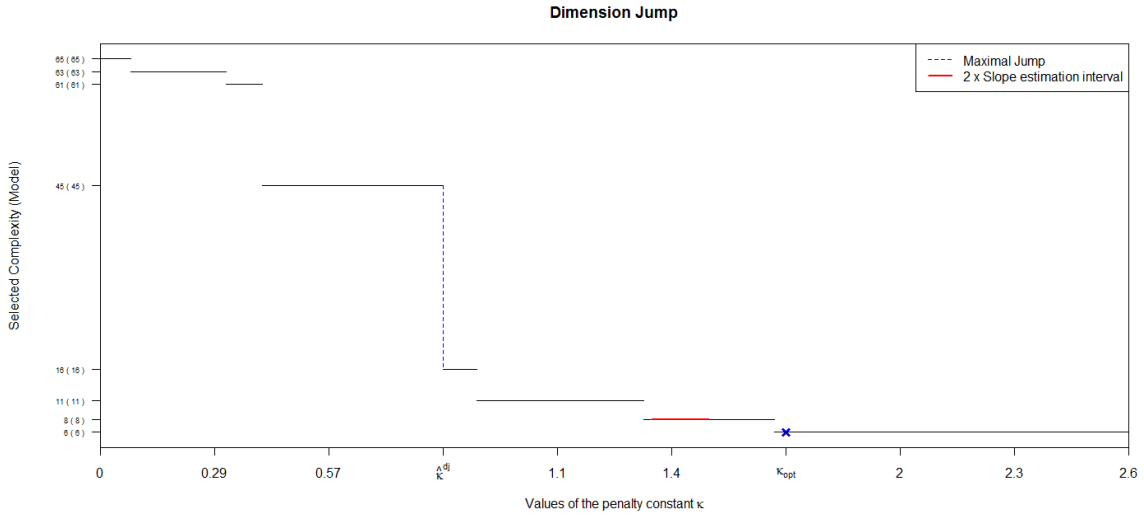


Figure 2.5: Illustration of dimension jump method $\beta = \beta_2$, $n = 1000$, $X = X^{(1)}$ with $\lambda_j = j^{-2}$ (*Capushe*).

we have $\text{pen}_{\text{opt}}(m) = \kappa_{\text{opt}}\sigma^2\frac{D_m}{n} \approx C - 2\gamma_n(\widehat{\beta}_m)$. Then, for m sufficiently large, the scatter plot of $(\sigma^2\frac{D_m}{n}, -\gamma_n(\widehat{\beta}_m))_m$ follows approximately a line of slope $\kappa_{\text{opt}}/2 = \kappa_{\min}$ which can be estimated by linear least-squares regression.

Figure 2.6 illustrates the slope estimation method. In particular, the left-hand plot represents the points $(\sigma^2\frac{D_m}{n}, -\gamma_n(\widehat{\beta}_m))_m$ and the fitted least squares line.

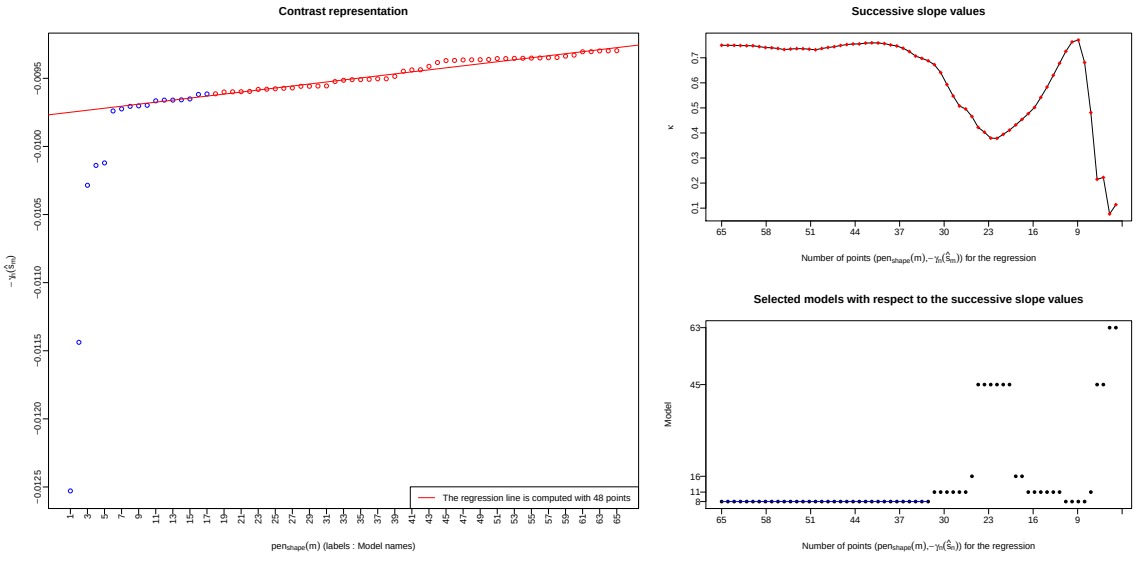


Figure 2.6: Illustration of the slope estimation method, $\beta = \beta_2$, $n = 1000$, $X = X^{(1)}$ with $\lambda_j = j^{-2}$ (*Capushe*).

Comparison of calibration by simulation and slope heuristics

We compare, in this section, the performances of the estimator $\widetilde{\beta}^{(FPCR)}$ when the constant κ is fixed, as recommended above, to $\kappa = 4$ or estimated by slope heuristics.

The results given in Table 2.1 suggest that the estimators are more stable when κ is fixed to $\kappa = 4$. The performance of the estimators obtained with the three methods (fixed constant ($\kappa = 4$), slope estimation and dimension jump) are almost identical when the sample size is large enough ($n = 1000$). The slope estimation method seems to behave better than the dimension jump method when the sample size is small but calibration by simulation behaves better than the two above for small sample sizes. From the considerations above, we choose to fix $\kappa = 4$ in the sequel.

λ_j		β_1			β_2		
		$n = 200$	$n = 500$	$n = 1000$	$n = 200$	$n = 500$	$n = 1000$
j^{-2}	$\kappa = 4$	**12.5 ±0.4**	**6.0 ±0.2**	**3.65 ±0.08**	**4.9 ± 0.3**	**1.90 ±0.09**	**0.91 ±0.05**
	DDSE	73 ±2	36 ±2	4.3 ±0.7	8 ±1	4.0 ±0.6	1.6 ±0.2
	Djump	100 ±20	11 ±5	**3.6 ±0.1**	6.2 ±0.4	2.6 ±0.2	1.3 ±0.1
j^{-3}	$\kappa = 4$	**6.9 ±0.3**	**3.2 ±0.1**	**1.83 ±0.05**	5.4 ±0.3	**1.8 ±0.2**	**0.84 ±0.04**
	DDSE	15.4 ±0.6	8.2 ±0.5	2.3 ±0.2	**5.0 ±0.3**	**2.0 ±0.2**	1.08 ±0.08
	Djump	51 ±4	13 ±3	1.89 ±0.09	6.5 ±0.5	2.7 ±0.3	1.4 ±0.2
e^{-j}	$\kappa = 4$	**4.9 ±0.3**	**2.02 ±0.09**	**1.08 ±0.05**	4.9 ±0.2	**1.9 ±0.1**	**0.79 ±0.05**
	DDSE	**5.1 ±0.3**	**2.1 ±0.1**	**1.11 ±0.05**	**4.5 ±0.2**	**2.0 ±0.1**	0.94 ±0.06
	Djump	53 ±3	27 ±3	4±2	7.2 ±0.9	3.7 ±0.7	1.6 ±0.5

Table 2.1: Mean prediction error and 95% confidence interval ($\times 10^{-4}$) over 500 Monte-Carlo replications of $\widetilde{\beta}^{(FPCR)}$. First line, dimension selected with the penalty (2.15), with $\kappa = 4$ and σ^2 known, DDSE: slope estimation method, Djump: dimension jump method.

2.4.4 Effect of the random term $\widehat{\sigma}^2_m$ in the penalty

We want to know what is the effect of the estimation of the noise σ^2 on the quality of estimation. We then compare the mean prediction error of the estimator $\widehat{\beta}^{(FPCR)}_{\widehat{m}^{(FPCR)}_\sigma}$ where

$$\widehat{m}^{(FPCR)}_\sigma \in \arg\min_{m=1,\dots,N_n} \gamma_n\left(\widehat{\beta}^{(FPCR)}_m\right) + \kappa\sigma^2\frac{D_m}{n},$$

with our estimator $\widehat{\beta}^{(FPCR)}_{\widehat{m}^{(FPCR)}}$, which is selected by minimization of

$$\gamma_n\left(\widehat{\beta}^{(FPCR)}_m\right) + \kappa\widehat{\sigma}^2_m\frac{D_m}{n},$$

where

$$\widehat{\sigma}^2_m = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \langle\widehat{\beta}^{(FPCR)}_m, X_i\rangle\right)^2.$$

We fix $\kappa = 4$ in both cases.

As we can see on the Table 2.2, the mean prediction error is quasi identical for both estimators $\widehat{\beta}^{(FPCR)}_{\widehat{m}^{(FPCR)}_\sigma}$ and $\widehat{\beta}^{(FPCR)}_{\widehat{m}^{(FPCR)}}$. Moreover, Table 2.3 shows that both criteria selects the same dimension in most cases. Then, even if the estimator $\widehat{\sigma}^2_m$, is not a good estimator of

λ_j		β_1			β_2		
		$n = 200$	$n = 500$	$n = 1000$	$n = 200$	$n = 500$	$n = 1000$
j^{-2}	$\widehat{m}_\sigma^{(FPCR)}$	12.6 ±0.4	6.1 ±0.2	3.65 ±0.08	5.0 ±0.3	2.0 ±0.1	0.93 ±0.05
	$\widehat{m}^{(FPCR)}$	12.3 ±0.4	6.0 ±0.2	3.59 ±0.08	5.1 ±0.3	2.0 ±0.1	0.93 ±0.05
j^{-3}	$\widehat{m}_\sigma^{(FPCR)}$	6.9 ±0.3	3.3 ±0.1	1.85 ±0.05	5.5 ±0.3	1.9 ±0.1	0.88 ±0.04
	$\widehat{m}^{(FPCR)}$	6.8 ±0.3	3.2 ±0.1	1.86 ±0.05	5.3 ±0.3	1.9 ±0.1	0.88 ±0.04
e^{-j}	$\widehat{m}_\sigma^{(FPCR)}$	4.9 ±0.2	2.05 ±0.09	1.11 ±0.05	4.9 ±0.2	2.0 ±0.1	0.86 ±0.05
	$\widehat{m}^{(FPCR)}$	4.9 ±0.2	2.05 ±0.09	1.10 ±0.05	4.9 ±0.2	2.0 ±0.1	0.86 ±0.05

Table 2.2: Mean prediction error and 95% confidence intervals ($\times 10^{-4}$) calculated on 500 estimators.

λ_j	β_1			β_2		
	$n = 200$	$n = 500$	$n = 1000$	$n = 200$	$n = 500$	$n = 1000$
j^{-2}	72	85	91	87	97	98
j^{-3}	84	91	94	91	97	98
e^{-j}	91	96	97	91	96	97

Table 2.3: Proportion of samples (in %) such that $\widehat{m}^{(FPCR)} = \widehat{m}_\sigma^{(FPCR)}$, calculated on 500 estimators.

σ^2 (especially when m is small), the criterion with random penalty behaves well.

2.4.5 Comparison of estimators $\widetilde{\beta}^{(KB)}$ and $\widetilde{\beta}^{(FPCR)}$

***Circular data* context**

To see the effect of estimation of eigenfunctions, we compare both estimators in a *circular data* context.

First remark that, with the choice of the Fourier basis $(\psi_j^{(2)})_{j\geq 1}$ the random function $X^{(2)}$ satisfies the assumptions of *circular* data ($X^{(2)}(0) = X^{(2)}(1)$ and $X^{(2)}$ is second-order stationary) as soon as J is odd and $\lambda_{2k} = \lambda_{2k+1}$ for all k.

The rank J must be chosen sufficiently large so that λ_J is negligible with respect to λ_1. We fix $J = 101$ in the following.

Figures 2.7 and 2.8 show that both estimators seem to be satisfactory. The boxplots of the last line indicate that the estimator $\widetilde{\beta}^{(KB)}$ has better performances than the estimator $\widetilde{\beta}^{(FPCR)}$ which was to be expected.

Non-circular data

In figures 2.9 and 2.10, we see that the FPCR estimator $\widetilde{\beta}^{(FPCR)}$ seems to have better performances than $\widetilde{\beta}^{(KB)}$. The dimension selection criterion, which is adapted to the context where the model is spanned by the eigenfunctions of the covariance operator (or their empirical counterpart), is not effective for the estimator $\widetilde{\beta}^{(KB)}$ – which is defined on the Fourier basis – when the curve X is not periodic. We see here the advantage of taking a data-driven basis: consistently with the theoretical results, the estimation procedure is efficient regardless the law of X.

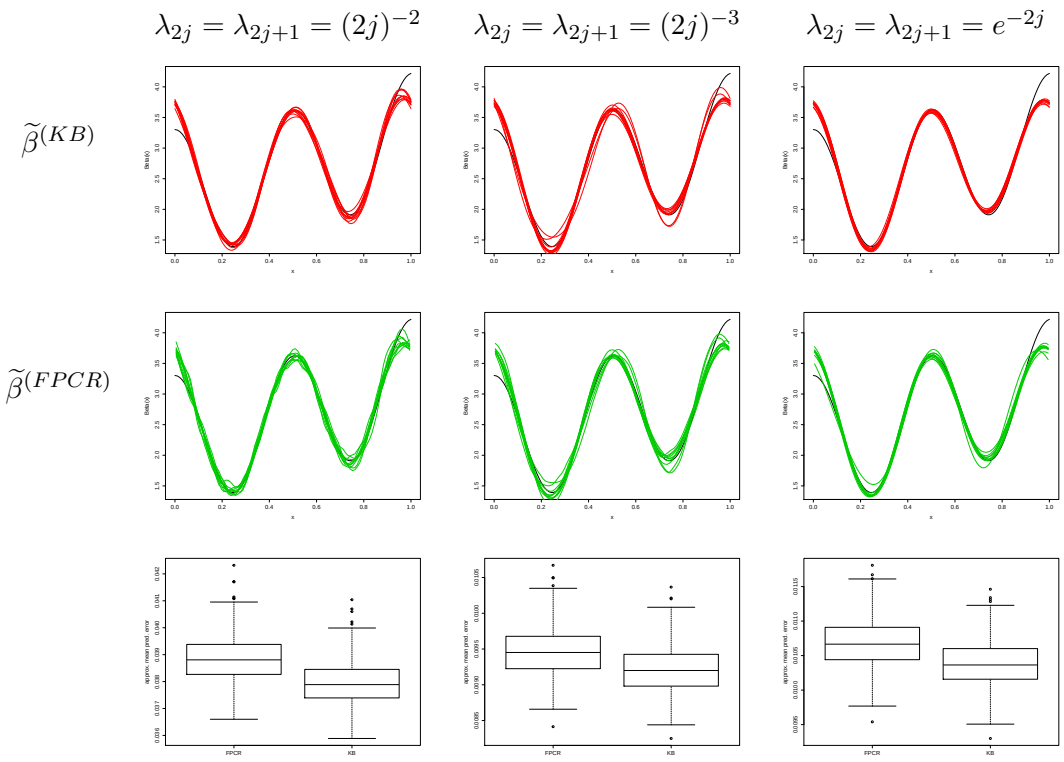


Figure 2.7: Estimation of β_1, $X = X^{(2)}$ (*circular* data). First line: plots of 10 independent realisations of $\widetilde{\beta}^{(KB)}$ (in red). Second line: plots of 10 independent realisations of $\widetilde{\beta}^{(FPCR)}$ (in green). In both lines β is plotted in black. Third line: comparison of mean squared prediction error over 1000 Monte-Carlo replications of $\widetilde{\beta}^{(KB)}$ and $\widetilde{\beta}^{(FPCR)}$. $n = 1000$.

2.4.6 Comparison with cross validation

We compare our dimension selection criterion with two cross validation criteria frequently used in practice. The first method consists in minimizing

$$GCV(m) := \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\left(1 - \mathrm{tr}(\mathbf{H}_m)/n\right)^2},$$

where $\widehat{Y}_i := \int_0^1 \widehat{\beta}_m(t) X_i(t) dt$ and $\mathbf{H}_m$ is the classical hat matrix defined by $\widehat{\mathbf{Y}} = (\widehat{Y}_1, ..., \widehat{Y}_n)' = \mathbf{H}_m \mathbf{Y}$. This criterion has been proposed in a similar context by Marx and Eilers (1996) and in the context of Functional Linear Models by Cardot, Ferraty, and Sarda (2003). The second one consists in minimizing the criterion

$$CV(m) := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i^{(-i)}\right)^2,$$

which has been proposed in the framework of Functional Linear Model by Hall and Hosseini-Nasab, 2006. Here $\widehat{Y}_i^{(-i)}$ is the value of Y_i predicted from the sample $\{(X_j, Y_j), j \neq i\}$. Note that an immediate drawback of this criterion is that it requires a much longer CPU time

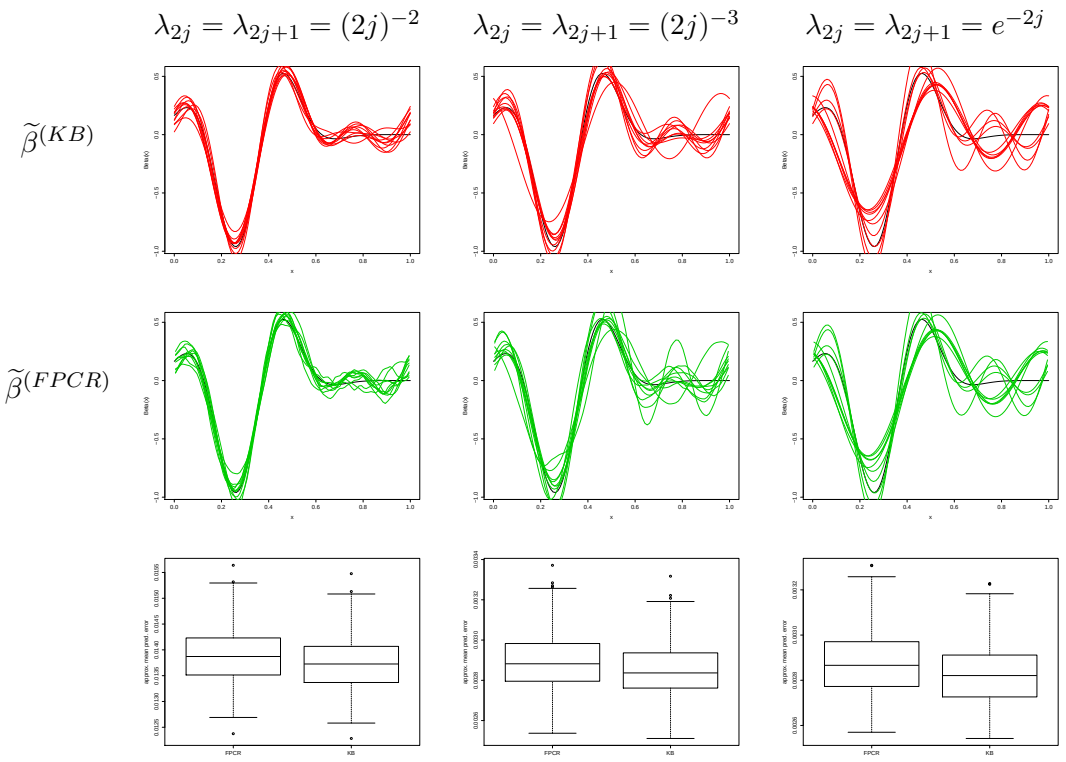


Figure 2.8: Estimation of β_2, $X = X^{(2)}$ (*circular* data). First line: plots of 10 independent realisations of $\widetilde{\beta}^{(KB)}$ (in red). Second line: plots of 10 independent realisations of $\widetilde{\beta}^{(FPCR)}$ (in green). In both lines β is plotted in black. Third line: comparison of mean squared prediction error over 1000 Monte-Carlo replications of $\widetilde{\beta}^{(KB)}$ and $\widetilde{\beta}^{(FPCR)}$. $n = 1000$.

than the GCV criterion or our penalized criterion.

λ_j		β_1			β_2		
j^{-2}		$n = 200$	$n = 500$	$n = 1000$	$n = 200$	$n = 500$	$n = 1000$
	crit	**12.5 ±0.4**	**5.8 ±0.2**	3.51 ±0.08	**4.7 ±0.3**	**1.89 ±0.09**	**0.89 ±0.05**
	GCV	80 ±2	55 ±2	47 ±2	80 ±2	55 ±2	47 ±2
	CV	**12.2 ±0.5**	**5.6 ±0.2**	**3.34 ±0.09**	5.7 ±0.4	2.2 ±0.2	1.08 ±0.06
j^{-3}	crit	**6.6 ±0.3**	**3.2 ±0.1**	**1.83 ±0.06**	**5.4 ±0.3**	**1.8 ±0.1**	**0.88 ±0.04**
	GCV	18.4 ±0.5	12.6 ±0.3	9.3 ±0.2	18.5 ±0.5	12.7 ±0.3	9.5 ±0.2
	CV	7.3 ±0.4	**3.3 ±0.2**	**1.78 ±0.07**	**5.1 ±0.4**	2.0 ±0.2	1.05 ±0.08
e^{-j}	crit	**4.7 ±0.2**	**2.03 ±0.08**	**1.11 ±0.05**	**4.9 ±0.2**	**1.82 ±0.09**	**0.77 ±0.05**
	GCV	5.8 ±0.3	2.67 ±0.09	1.45 ±0.05	6.0 ±0.3	2.7 ±0.1	1.40 ±0.05
	CV	**4.8 ±0.3**	**2.05 ±0.09**	**1.10 ±0.05**	**4.6 ±0.3**	**1.8 ±0.1**	0.87 ±0.05

Table 2.4: Mean prediction error ($\times 10^{-4}$) and approximated 95% confidence interval (calculated from 500 independent samples of size $n = 1000$).

According to Figure 2.11 and Table 2.4, performances of our estimator seem to be quite similar to the functional PCR estimator with dimension selected by minimization of the CV criterion. Conversely, the GCV criterion selects systematically the highest dimensional model which leads to poor performances.

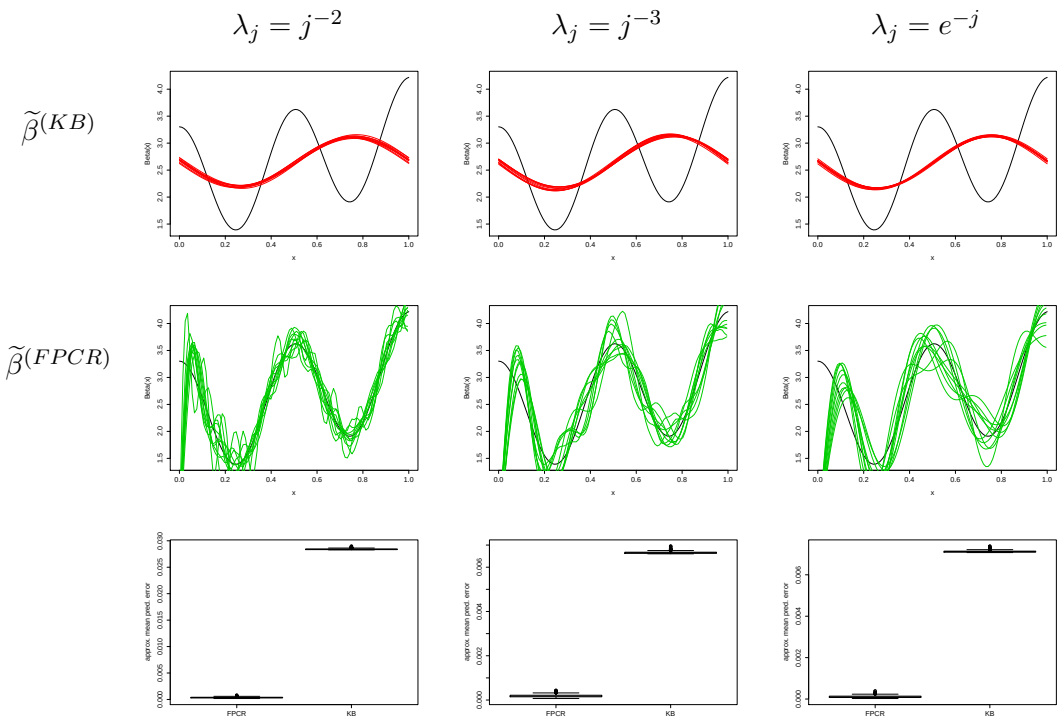


Figure 2.9: Estimation of β_1, $X = X^{(1)}$. First line: plots of 10 independent realisations of $\widetilde{\beta}^{(KB)}$ (in red). Second line: plots of 10 independent realisations of $\widetilde{\beta}^{(FPCR)}$ (in green). In both lines β is plotted in black. Third line: comparison of mean squared prediction error over 1000 Monte-Carlo replications of $\widetilde{\beta}^{(KB)}$ and $\widetilde{\beta}^{(FPCR)}$. $n = 1000$.

2.4.7 Comparison with the pseudo-oracle

The purpose of this section is to compare the different methods of dimension selection previously considered with the pseudo-oracle dimension.

We first compare the distance to the oracle of our method with slope heuristics methods. The histogram in Figure 2.12 suggests that the slope estimation methods and the jump dimension method tend to select higher dimensions – that is to say to under-penalize – and have the disadvantage of being slightly unstable. Then we compare with cross-validated criteria. The criterion CV seems to have a behaviour very similar to slope heuristic methods, as shown in Figure 2.13. Conversely, setting $\kappa = 4$, the dimension selected by our criterion is usually slightly lower than the pseudo-oracle, which means that it over-penalizes. As shown in Figure 2.4, this problem can be solved by choosing a smaller value of κ (e.g. $\kappa = 2$) but at the price of the stability of the estimator.

We also look at the ratio to the pseudo-oracle that is to say the quantity

$$\mathbb{E}\left[\frac{\|\beta - \widehat{\beta}_{\widehat{m}}\|_\Gamma^2}{\|\beta - \widehat{\beta}_{m^*}\|_\Gamma^2}\right],$$

the results are given in Table 2.5 and confirm findings discussed above.

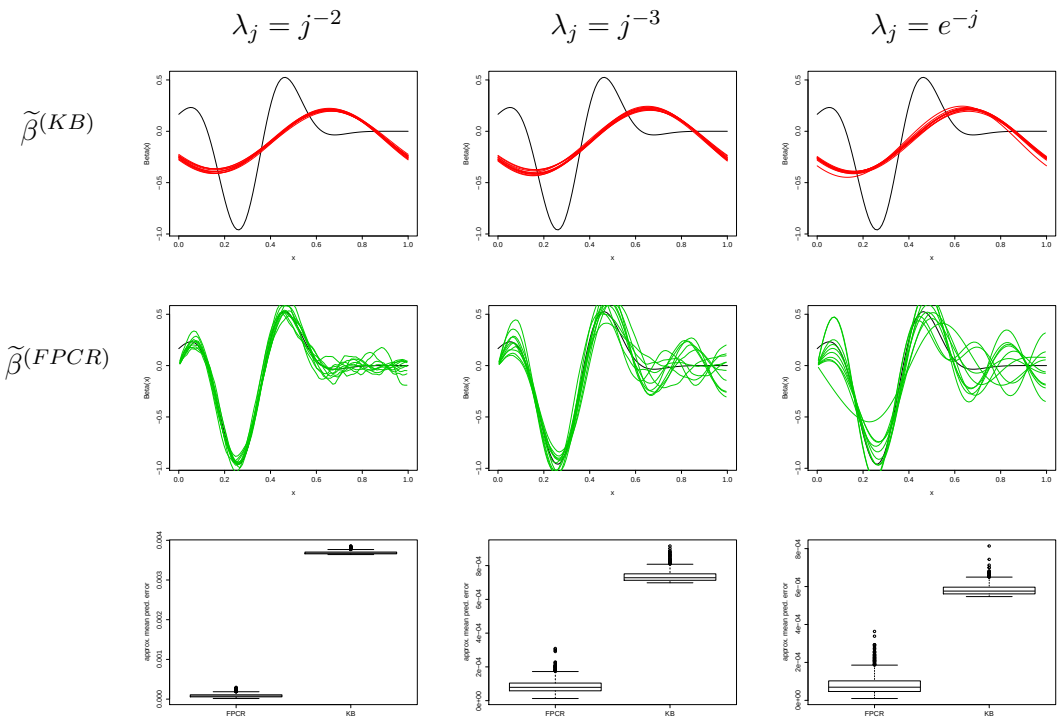


Figure 2.10: Estimation of β_2, $X = X^{(2)}$. First line: plots of 10 independent realisations of $\widetilde{\beta}^{(KB)}$ (in red). Second line: plots of 10 independent realisations of $\widetilde{\beta}^{(FPCR)}$ (in green). In both lines β is plotted in black. Third line: comparison of mean squared prediction error over 1000 Monte-Carlo replications of $\widetilde{\beta}^{(KB)}$ and $\widetilde{\beta}^{(FPCR)}$. $n = 1000$.

λ_j		β_1			β_2		
		$n = 200$	$n = 500$	$n = 1000$	$n = 200$	$n = 500$	$n = 1000$
j^{-2}	crit	**1.43 ±0.04**	1.33 ±0.03	1.33 ±0.03	**1.32 ±0.06**	**1.27 ±0.05**	**1.13 ±0.03**
	GCV	9.7 ±0.2	13.0 ±0.4	19.6 ±0.7	28 ±2	47 ±3	76 ±5
	CV	**1.40 ±0.06**	**1.27 ±0.03**	**1.25 ±0.03**	1.6 ±0.2	1.51 ±0.09	1.44 ±0.08
	DDSE	1.7 ±0.1	1.5 ±0.1	1.45 ±0.09	2.3 ±0.4	2.2 ±0.4	2.4 ±0.6
	Djump	1.48 ±0.06	1.34 ±0.04	1.30 ±0.04	1.7 ±0.2	1.7 ±0.2	1.6 ±0.2
j^{-3}	crit	**1.44 ±0.05**	**1.42 ±0.04**	1.45 ±0.04	2.1 ±0.2	**1.41 ±0.07**	**1.33 ±0.06**
	GCV	4.3 ±0.2	5.9 ±0.3	7.9 ±0.3	7.8 ±0.6	12.3 ±0.8	18 ±2
	CV	1.56 ±0.08	**1.43 ±0.06**	**1.39 ±0.05**	**1.8 ±0.2**	1.7 ±0.2	1.6 ±0.2
	DDSE	2.0 ±0.2	1.8 ±0.1	1.7 ±0.2	2.7 ±0.3	2.5 ±0.3	2.4 ±0.3
	Djump	1.51 ±0.06	1.47 ±0.05	1.46 ±0.06	**1.9 ±0.2**	1.8 ±0.2	1.8 ±0.2
e^{-j}	crit	1.43 ±0.06	**1.37 ±0.04**	**1.40 ±0.05**	**2.0 ±0.2**	**1.7 ±0.1**	**1.26 ±0.05**
	GCV	1.80 ±0.07	1.90 ±0.08	1.91 ±0.07	2.4 ±0.2	2.7 ±0.2	2.7 ±0.2
	CV	1.42 ±0.06	**1.40 ±0.07**	**1.40 ±0.06**	**1.8 ±0.2**	**1.6 ±0.2**	1.5 ±0.1
	DDSE	**1.34 ±0.04**	**1.41 ±0.06**	**1.43 ±0.05**	**1.8 ±0.2**	**1.7 ±0.2**	1.66 ±0.08
	Djump	8 ±2	6 ±2	6 ±3	2.7 ±0.4	3 ±1	3 ±2

Table 2.5: Mean ratio to the pseudo-oracle and 95% confidence interval calculated from 500 samples.

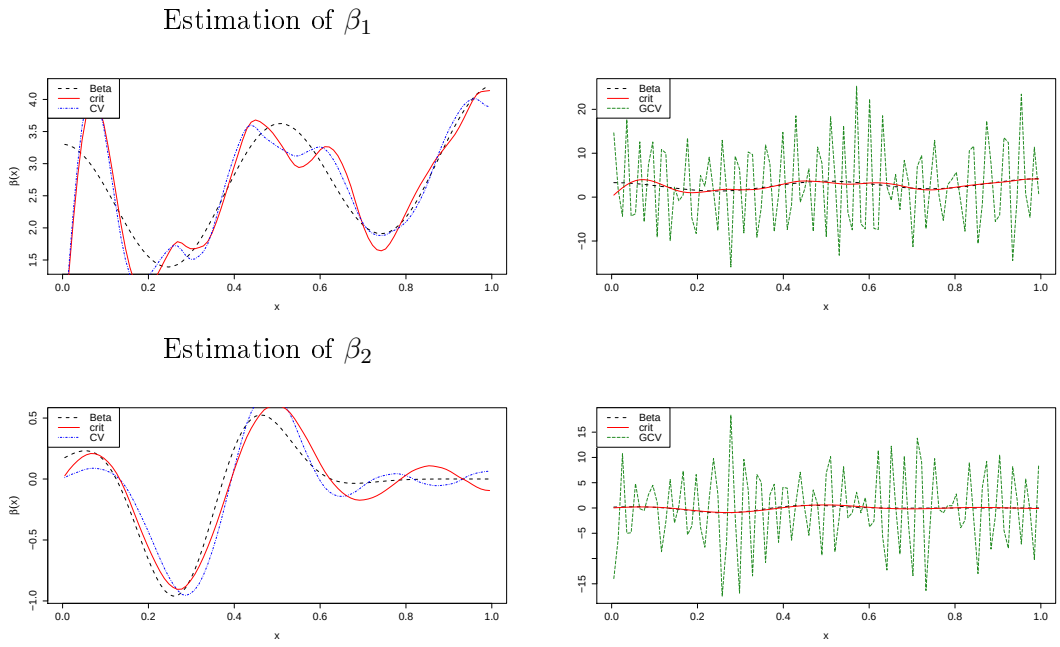


Figure 2.11: Left: comparison of estimators $\widehat{\beta}_m$ when m is selected by minimization of our penalized criterion or the CV criterion. Right: comparison with the GCV criterion. $n = 2000$, $\lambda_j = j^{-3}$.

2.4.8 Addition of a noise on the covariate X

In this section, we seek to observe how the quality of the estimate degrades by the addition of a noise on the covariate X. We do not calculate the estimator with the sample $\{(X_i, Y_i), i = 1, ..., n\}$ as above but with the sample $\{(W_i, Y_i), i = 1, ..., n\}$ where W_i is defined by

$$W_i(t_j) := X_i(t_j) + \delta_{i,j},$$

where $(t_j)_{j=0}^p = \{(j-1)/(p-1)\}_{j=1}^p$ and $\{\delta_{i,j}, i = 1, ..., n, j = 1, ..., p\}$ is an i.i.d. sequence of centred normal random variables with variance

$$\mathrm{Var}(\delta_{i,j}) = \sigma_\delta^2.$$

This definition corresponds to that given by Li and Hsing, 2007 and also by Crambes, Kneip, and Sarda, 2009 which have adapted their estimator in the specific case of noisy data from an estimator of the variance σ_δ^2. This new estimator reaches the same convergence rate as the one proposed when the data is not noisy provided p is quite large compared to n. They compared the two estimators on a real dataset, their results show a slight advantage for the estimator taking into account the measurement error on the $X_i(t_j)$.

Here we keep the same estimator and the aim is to study its behaviour for different noise levels. Figure 2.14 represents the effect of adding different levels of noise on a realisation of the variable X. On Figure 2.15, we can see that, for the smallest noise levels ($\sigma_\delta^2 = 0.001$ and $\sigma_\delta^2 = 0.01$), the performances of the estimator are very close, which suggests that the

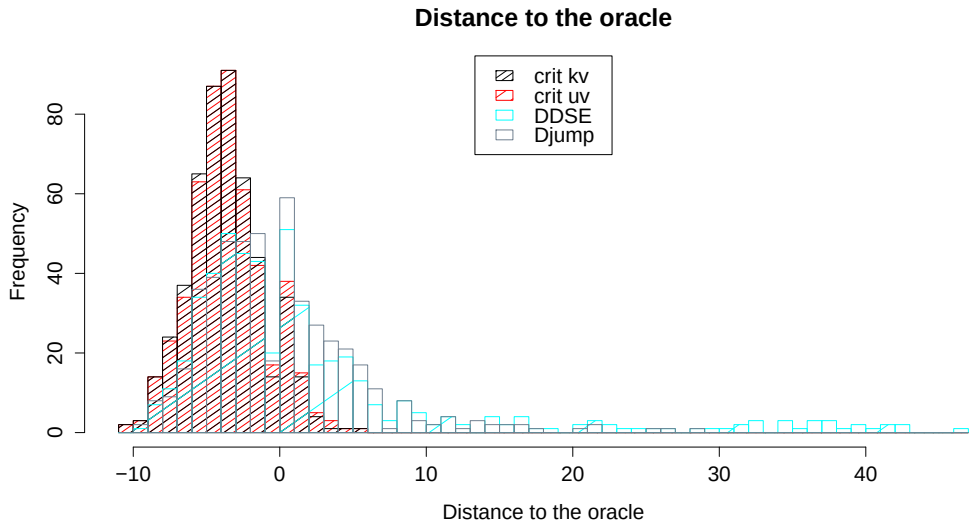


Figure 2.12: Histogram of distances to the oracle calculated from 500 independent samples of size $n = 2000$ with our criterion when $\kappa = 4$ with known noise variance σ^2 (crit kv) or unknown noise variance (crit uv) or when κ is calibrated with the slope estimation method (DDSE) or dimension jump method (Djump) (β_1, $\lambda = j^{-3}$).

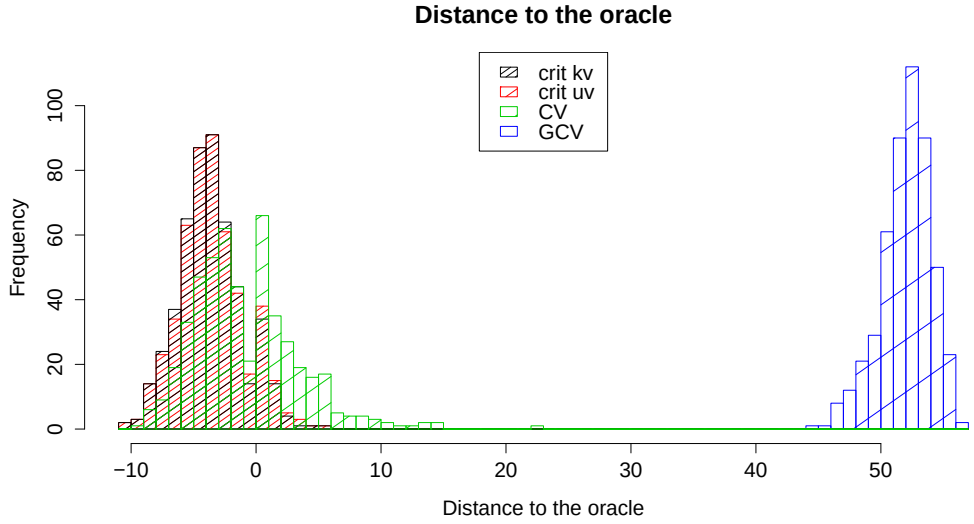


Figure 2.13: Histogram of distances to the oracle calculated from 500 independent samples of size $n = 2000$ with our criterion with known variance (crit kv) or unknown variance (crit uv) and with cross-validated methods (β_1, $\lambda = j^{-3}$).

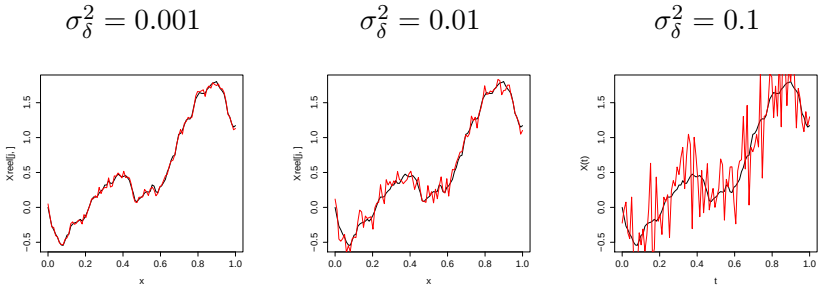


Figure 2.14: Effect of different levels of noise. Black: a realisation of the variable X, red: a realisation of the corresponding variable W.

estimation is very stable under small perturbations of X. These results are confirmed by the Monte-Carlo study presented in Figure 2.16.

However, it could be interesting to complete this numerical study by theoretical results. An adaptation of the estimate in presence of strong noise (Figure 2.14 – right), as proposed by Crambes, Kneip, and Sarda (2009), could also improve significantly the quality of estimation.

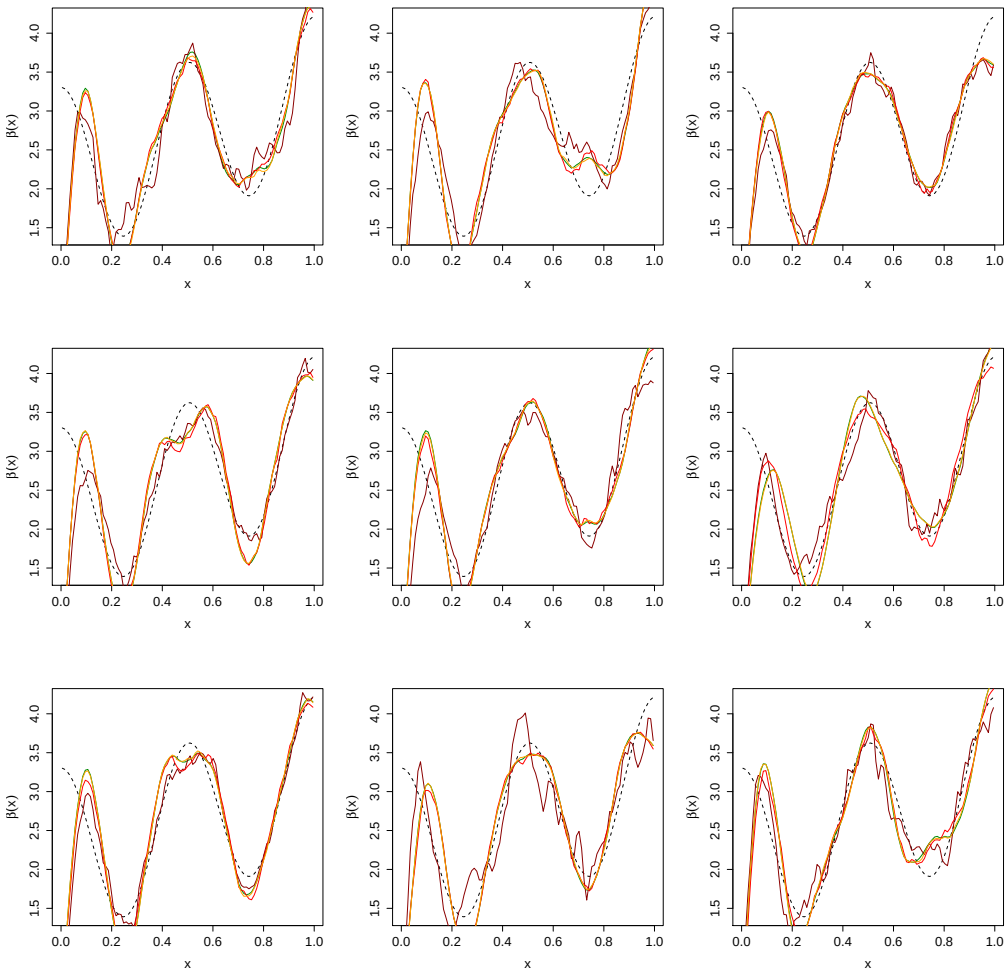


Figure 2.15: Estimator calculated under different noise levels ($n = 1000$): dotted line : function to estimate, green line: estimator calculated from non-noisy data, orange line: estimator calculated when : $\sigma_\delta^2 = 0.001$, red: $\sigma_\delta^2 = 0.01$, maroon: $\sigma_\delta^2 = 0.1$.

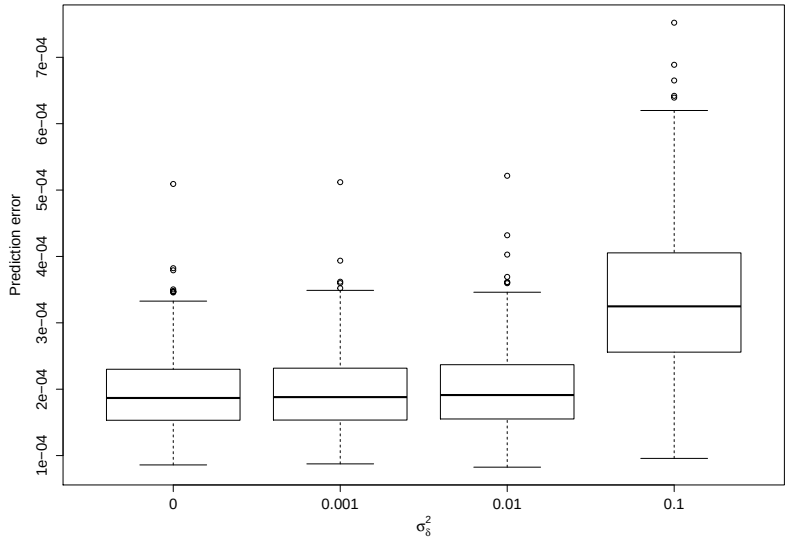


Figure 2.16: Boxplot of prediction errors for 500 Monte-Carlo realisations of $\widetilde{\beta}^{(FPCR)}$ under different noise levels, estimation of β_1, $n = 1000$.

2.5 Proofs

2.5.1 Control of the random penalty

We give here a first lemma allowing to control the randomness in the definition of the penalty.

Lemma 1. *Under Assumption **H1**, set $\kappa := \theta(1+\delta)$, we have, for all $m \in \widehat{\mathcal{M}_n}$, when $\widetilde{\beta} = \widetilde{\beta}^{(FPCR)}$,*

$$\mathbb{E}_{\mathbf{X}}\left[(\widehat{\text{pen}}(m) - \text{pen}(m))\right] \leq \kappa \frac{D_m}{n} \|\beta - \widehat{\Pi}_m \beta\|^2_{\Gamma_n}, \tag{2.16}$$

where $\mathbb{E}_{\mathbf{X}}$ is the conditional expectation with respect to $\mathbf{X} = (X_1, \ldots, X_n)$ and

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}[(\text{pen}(\widehat{m}^{(FPCR)}) - \widehat{\text{pen}}(\widehat{m}^{(FPCR)}))] \quad &\leq \quad \frac{\kappa}{n}\mathbb{E}_{\mathbf{X}}[2D_{\widehat{m}^{(FPCR)}}\nu_n(\widehat{\Pi}_m\beta - \widetilde{\beta}^{(FPCR)})] \\ &\quad + \frac{\kappa D_{\widehat{N}_n^{(FPCR)}}}{n\sqrt{n}}\left(\sqrt{\text{Var}(\varepsilon^2)} + 2\|\beta\|_\Gamma \sigma\right), \end{aligned} \tag{2.17}$$

with $\nu_n : t \in \mathbb{H} \mapsto \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle t, X_i\rangle$.
The same inequalities hold for $\widetilde{\beta} = \widetilde{\beta}^{(KB)}$, replacing $\widehat{\Pi}_m$ by Π_m and $\widehat{\mathcal{M}_n}$ by $\mathcal{M}_n$.

Proof. By definitions of $\widehat{\sigma}^2_m = \gamma_n(\widehat{\beta}^{(FPCR)}_m)$ and $\widehat{\beta}^{(FPCR)}_m$, we have $\widehat{\sigma}^2_m \leq \gamma_n(\widehat{\Pi}_m\beta)$, then

$$\mathbb{E}_{\mathbf{X}}[\widehat{\text{pen}}(m) - \text{pen}(m)] = \kappa\frac{D_m}{n}\mathbb{E}_{\mathbf{X}}[\widehat{\sigma}^2_m - \sigma^2] \leq \kappa\frac{D_m}{n}\mathbb{E}_{\mathbf{X}}[\gamma_n(\widehat{\Pi}_m\beta) - \sigma^2].$$

Now, by independence of ε_i with $< \beta - \widehat{\Pi}_m\beta, X_i >$ and by definition of $Y_i = \langle \beta, X_i \rangle + \varepsilon$,

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\gamma_n(\widehat{\Pi}_m\beta) - \sigma^2] &= \mathbb{E}_{\mathbf{X}}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - < \widehat{\Pi}_m\beta, X_i >)^2 - \sigma^2\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 - 2\varepsilon_i < \beta - \widehat{\Pi}_m\beta, X_i > + < \beta - \widehat{\Pi}_m\beta, X_i >^2 - \sigma^2\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\frac{1}{n}\sum_{i=1}^{n} < \beta - \widehat{\Pi}_m\beta, X_i >^2\right] = \mathbb{E}_{\mathbf{X}}[\|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2] \\
&= \|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2,
\end{aligned}$$

and Equation (2.16) follows.

Likewise, set $\widetilde{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$, since $\widehat{\sigma}^2_{\widehat{m}^{(FPCR)}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle \widetilde{\beta}^{(FPCR)}, X_i \rangle)^2$,

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\mathrm{pen}(\widehat{m}^{(FPCR)}) - \widehat{\mathrm{pen}}(\widehat{m}^{(FPCR)})] &= \frac{\kappa}{n}\mathbb{E}_{\mathbf{X}}[D_{\widehat{m}^{(FPCR)}}(\sigma^2 - \widehat{\sigma}^2_{\widehat{m}^{(FPCR)}})] \\
&= \frac{\kappa}{n}\left(\mathbb{E}_{\mathbf{X}}[D_{\widehat{m}^{(FPCR)}}(\sigma^2 - \widetilde{\sigma}^2)] - \mathbb{E}_{\mathbf{X}}\left[\frac{D_{\widehat{m}^{(FPCR)}}}{n}\sum_{i=1}^{n}\langle \beta - \widetilde{\beta}^{(FPCR)}, X_i\rangle^2\right]\right. \\
&\qquad \left. + \mathbb{E}_{\mathbf{X}}\left[\frac{2D_{\widehat{m}^{(FPCR)}}}{n}\sum_{i=1}^{n}\varepsilon_i\langle \beta - \widetilde{\beta}^{(FPCR)}, X_i\rangle\right]\right) \\
&\leq \frac{\kappa}{n}\left(\mathbb{E}_{\mathbf{X}}[D_{\widehat{m}^{(FPCR)}}(\sigma^2 - \widetilde{\sigma}^2)] + 2\mathbb{E}_{\mathbf{X}}\left[D_{\widehat{m}^{(FPCR)}}\nu_n\left(\beta - \widetilde{\beta}^{(FPCR)}\right)\right]\right) \\
&\leq \frac{\kappa}{n}\left(\mathbb{E}_{\mathbf{X}}[D_{\widehat{m}^{(FPCR)}}(\sigma^2 - \widetilde{\sigma}^2)] + 2\mathbb{E}_{\mathbf{X}}\left[D_{\widehat{m}^{(FPCR)}}\nu_n(\beta - \widehat{\Pi}_m\beta)\right]\right) \\
&\qquad + \frac{\kappa}{n}\mathbb{E}_{\mathbf{X}}\left[D_{\widehat{m}^{(FPCR)}}\nu_n\left(\widehat{\Pi}_m\beta - \widetilde{\beta}^{(FPCR)}\right)\right],
\end{aligned}$$

where $\nu_n : t \mapsto \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle t, X_i\rangle$. By Cauchy-Schwarz's Inequality, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[D_{\widehat{m}^{(FPCR)}}(\sigma^2 - \widetilde{\sigma}^2)] &\leq D_{\widehat{N}_n^{(FPCR)}}\mathbb{E}_{\mathbf{X}}[(\sigma^2 - \widetilde{\sigma}^2)^2]^{1/2} \\
&= D_{\widehat{N}_n^{(FPCR)}}\left(\frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}_{\mathbf{X}}[(\varepsilon_i^2 - \sigma^2)(\varepsilon_j^2 - \sigma^2)]\right)^{1/2} \\
&= D_{\widehat{N}_n^{(FPCR)}}\sqrt{\frac{\mathrm{Var}(\varepsilon^2)}{n}}.
\end{aligned}$$

Next, since the ε_i's are independent of the X_i's and by consequence of $\widehat{\Pi}_m$, we have:

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[D_{\widehat{m}^{(FPCR)}}\nu_n(\beta - \widehat{\Pi}_m\beta)] &\leq D_{\widehat{N}_n^{(FPCR)}}\mathbb{E}_{\mathbf{X}}[\nu_n^2(\beta - \widehat{\Pi}_m\beta)]^{1/2} \\
&\leq \frac{D_{\widehat{N}_n^{(FPCR)}}}{n}\left(\sum_{i_1,i_2=1}^{n}\mathbb{E}_{\mathbf{X}}[\varepsilon_{i_1}\varepsilon_{i_2} < \beta - \widehat{\Pi}_m\beta, X_{i_1} >< \beta - \widehat{\Pi}_m\beta, X_{i_2} >]\right)^{1/2} \\
&\leq \frac{D_{\widehat{N}_n^{(FPCR)}}}{\sqrt{n}}\sigma\mathbb{E}[\|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2]^{1/2} \leq \frac{D_{\widehat{N}_n^{(FPCR)}}}{\sqrt{n}}\sigma\|\beta\|_{\Gamma},
\end{aligned}$$

since $D_{\widehat{N}_n^{(FPCR)}}$ depends only on $\mathbf{X}$. □

2.5.2 Proof of Proposition 1

Proof. We start here with the proof of Inequality (2.9). We provide in a first step an oracle-inequality conditionally to $\mathbf{X} = \{X_1, \ldots, X_n\}$ allowing to use classical model selection tools. We follow mainly the sketch of proof of Baraud (2000, Corollary 3.1) for regression on a fixed design with adaptation in order to take into account the randomness of both penalty and dimension collection $\widehat{\mathcal{M}_n}$ (for $\widetilde{\beta}^{(FPCR)}$).

We first prove Inequality (2.9).

The following proof is based on contrast decomposition and the control of the remaining empirical process. More precisely, by definition of $\widehat{m}$, for all $m \in \widehat{\mathcal{M}}_n$

$$\gamma_n(\widetilde{\beta}^{(FPCR)}) + \widehat{\text{pen}}(\widehat{m}) \leq \gamma_n(\widehat{\beta}_m^{(FPCR)}) + \widehat{\text{pen}}(m)$$

and by definition of $\widehat{\beta}_m^{(FPCR)}$:

$$\gamma_n(\widehat{\beta}_m^{(FPCR)}) \leq \gamma_n(\widehat{\Pi}_m \beta),$$

then

$$\gamma_n(\widetilde{\beta}^{(FPCR)}) - \gamma_n(\widehat{\Pi}_m \beta) \leq \widehat{\text{pen}}(m) - \widehat{\text{pen}}(\widehat{m}).$$

Moreover

$$\gamma_n(\widetilde{\beta}^{(FPCR)}) - \gamma_n(\widehat{\Pi}_m \beta) = \|\beta - \widetilde{\beta}^{(FPCR)}\|_{\Gamma_n}^2 - \|\beta - \widehat{\Pi}_m \beta\|_{\Gamma_n}^2 + 2\nu_n\left(\widehat{\Pi}_m \beta - \widetilde{\beta}^{(FPCR)}\right),$$

with

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i < t, X_i > .$$

Then,

$$\|\beta - \widetilde{\beta}^{(FPCR)}\|_{\Gamma_n}^2 \leq \|\beta - \widehat{\Pi}_m \beta\|_{\Gamma_n}^2 + \widehat{\text{pen}}(m) - \widehat{\text{pen}}(\widehat{m}) + 2\nu_n(\widetilde{\beta}^{(FPCR)} - \widehat{\Pi}_m \beta). \quad (2.18)$$

The first step is to replace the random function $\widehat{\text{pen}}$ by its empirical counterpart pen, this can be done by using the results of Lemma 1 p.56 directly in Equation (2.18):

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}[\|\beta - \widetilde{\beta}^{(FPCR)}\|_{\Gamma_n}^2] \quad &\leq \quad \left(1 + \kappa\frac{D_m}{n}\right)\|\beta - \widehat{\Pi}_m \beta\|_{\Gamma_n}^2 + \mathbb{E}_{\mathbf{X}}[\text{pen}(m) - \text{pen}(\widehat{m})] \\ +\mathbb{E}_{\mathbf{X}}\left[2\left(1 + \frac{\kappa D_{\widehat{m}}}{n}\right)\nu_n(\widetilde{\beta}^{(FPCR)} - \widehat{\Pi}_m \beta)\right] &+ \kappa\frac{D_{\widehat{N}_n^{(FPCR)}}}{n\sqrt{n}}\left(\sqrt{\text{Var}(\varepsilon^2)} + 2\|\beta\|_\Gamma \sigma\right).\end{aligned}$$

Now, remarking that, since $n \geq 2$,

$$D_{\widehat{N}_n^{(FPCR)}} \leq 20\sqrt{n/\ln^3 n} \leq \left(20/\sqrt{\ln^{3/2} 2}\right)\sqrt{n},$$

and, in addition by Assumption **H1** and Hölder's inequality $\text{Var}(\varepsilon^2) = \mathbb{E}[\varepsilon^4] \leq \mathbb{E}[\varepsilon^p]^{2/p} =$

$\tau_p^{2/p}$ and $2\|\beta\|_\Gamma\sigma \leq \|\beta\|_\Gamma^2 + \sigma^2$. This gives us the following result

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}[\|\beta - \widetilde{\beta}^{(FPCR)}\|_{\Gamma_n}^2] &\leq \left(1 + \kappa\frac{D_m}{n}\right)\|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2 + \mathbb{E}_{\mathbf{X}}[\mathrm{pen}(m) - \mathrm{pen}(\widehat{m})] \\ &+\mathbb{E}_{\mathbf{X}}\left[2\left(1 + \frac{\kappa D_{\widehat{m}}}{n}\right)\nu_n(\widetilde{\beta}^{(FPCR)} - \widehat{\Pi}_m\beta)\right] + C\frac{\|\beta\|_\Gamma^2 + \tau_p^{2/p}}{n}, \qquad (2.19)\end{aligned}$$

where $\kappa := \theta(1+\delta)$ and $C > 0$ is a universal constant.

Then the last step consists in controlling the empirical linear process ν_n on $\widehat{S}_{m\vee\widehat{m}}$. Remark that for all $\delta > 0$, for all $m \in \widehat{\mathcal{M}_n}$,

$$2\nu_n(\widetilde{\beta}^{(FPCR)} - \widehat{\Pi}_m\beta) \leq \frac{1}{\theta}\|\widetilde{\beta}^{(FPCR)} - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2 + \theta \sup_{\substack{f\in\widehat{S}_{\widehat{m}\vee m}\\ \|f\|_{\Gamma_n}=1}} \nu_n^2(f), \qquad (2.20)$$

since for all $x, y \in \mathbb{R}$ and $\theta > 0$, $2xy \leq \theta^{-1}x^2 + \theta y^2$.
Let $p(m,m') := (1+\delta)\frac{D_{m\vee m'}}{n}\sigma^2$, remark that $\mathrm{pen}(m) + \mathrm{pen}(m') \geq \theta p(m,m')$. Then, since $\theta > 4$ and $\widehat{N}_n \leq n/\kappa$, gathering equations (2.19) and (2.20) we obtain:

$$\begin{aligned}\left(1 - \frac{4}{\theta}\right)\mathbb{E}_{\mathbf{X}}\left[\|\widetilde{\beta}^{(FPCR)} - \beta\|_{\Gamma_n}^2\right] &\leq \left(2 + \frac{4}{\theta}\right)\|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2 \\ &+2\mathrm{pen}(m) + 2\theta\mathbb{E}_{\mathbf{X}}\left[\left(\sup_{\substack{f\in\widehat{S}_{\widehat{m}\vee m}\\ \|f\|_{\Gamma_n}=1}} \nu_n^2(f) - p(m,\widehat{m})\right)_+\right].\end{aligned}$$

Then the last step is to bound the variations of $\sup_{\substack{f\in\widehat{S}_{\widehat{m}\vee m}\\ \|f\|_{\Gamma_n}=1}} \nu_n^2(f)$ (which can be seen as a variance term) around $p(m,m')$ and the result comes from Lemma 2 detailed below and the fact that $\sigma^2 = \mathbb{E}[\varepsilon^2] \leq \mathbb{E}[\varepsilon^p]^{2/p} = \tau_p^{2/p}$:

$$\mathbb{E}_{\mathbf{X}}\left[\|\widetilde{\beta}^{(FPCR)} - \beta\|_{\Gamma_n}^2\right] \leq C\left(\min_{m\in\widehat{\mathcal{M}_n}}\left\{\|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2 + \mathrm{pen}(m)\right\} + \frac{\|\beta\|_\Gamma^2 + \tau_p^{2/p}}{n}\right), \qquad (2.21)$$

where $C > 0$ depends only on θ, p and δ.

Now, we must replace $\widehat{\mathcal{M}_n}$ by its non random counterpart $\mathcal{M}_n$. First, we have

$$\begin{aligned}&\mathbb{E}_{\mathbf{X}}\left[\|\widetilde{\beta}^{(FPCR)} - \beta\|_{\Gamma_n}^2\right]\mathbf{1}_{\{N_n\leq\widehat{N}_n^{(FPCR)}\}} \\ &\qquad\leq C\left(\min_{m\in\mathcal{M}_n}\left\{\|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2 + \mathrm{pen}(m)\right\} + \frac{\|\beta\|_\Gamma^2 + \tau_p^{2/p}}{n}\right),\end{aligned}$$

since $\mathcal{M}_n \subset \widehat{\mathcal{M}_n}$ if $N_n \leq \widehat{N}_n^{(FPCR)}$. Now, the case $N_n > \widehat{N}_n^{(FPCR)}$ deserves further

attention. In that case, there exists $m \in \mathcal{M}_n$ such that $m > \widehat{N}_n^{(FPCR)}$. However

$$\begin{aligned}\|\beta - \widehat{\Pi}_{\widehat{N}_n^{(FPCR)}}\beta\|_{\Gamma_n}^2 + \mathrm{pen}(\widehat{N}_n^{(FPCR)}) &\leq \sum_{j>D_{\widehat{N}_n^{(FPCR)}}} \widehat{\lambda}_j < \beta, \widehat{\psi}_j >^2 + \mathrm{pen}(m) \\ &\quad + \frac{\theta(1+\delta)\sigma^2}{n}(D_{\widehat{N_n}^{(FPCR)}} - D_m) \\ &\leq \sum_{j>D_{\widehat{N}_n^{(FPCR)}}} \widehat{\lambda}_j < \beta, \widehat{\psi}_j >^2 + \mathrm{pen}(m),\end{aligned}$$

and

$$\begin{aligned}\sum_{j>D_{\widehat{N}_n^{(FPCR)}}} \widehat{\lambda}_j < \beta, \widehat{\psi}_j >^2 &= \|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2 + \sum_{j=D_{\widehat{N}_n^{(FPCR)}}+1}^{D_m} \widehat{\lambda}_j < \beta, \widehat{\psi}_j >^2 \\ &\leq \|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2 + \mathfrak{s}_n\|\beta\|^2.\end{aligned}$$

The last inequality comes from the fact that, for all $j \in \{D_{\widehat{N}_n^{(FPCR)}}+1, \ldots, D_{N_n}\}$, we have, by definition of N_n and since (D_m) is an increasing sequence, $j \leq \min\{20\sqrt{n/\ln^3(n)}, n/(\theta(1+2\delta))\}$. Hence, by definition of $\widehat{N}_n^{(FPCR)}$, $\widehat{\lambda}_j \leq \mathfrak{s}_n$. Therefore, as $\mathfrak{s}_n = \frac{2}{n^2}\left(1 - 1/\ln^2 n\right) \leq 2/n^2$, we get

$$\|\beta - \widehat{\Pi}_{\widehat{N}_n^{(FPCR)}}\beta\|_{\Gamma_n}^2 + \mathrm{pen}(\widehat{N}_n^{(FPCR)}) \leq \|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2 + \mathrm{pen}(m) + \frac{2\|\beta\|^2}{n^2}, \tag{2.22}$$

and we obtain, for all $m \in \mathcal{M}_n$, since $\|\beta\|_\Gamma^2 \leq \rho(\Gamma)\|\beta\|^2$ (where $\rho(\Gamma)$ is the *spectral radius* of Γ that is, in our context, its maximal eigenvalue):

$$\mathbb{E}_{\mathbf{X}}\left[\|\beta - \widehat{\beta}_{\widehat{m}}^{(FPCR)}\|_{\Gamma_n}^2\right] \leq C\left(\|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2 + \mathrm{pen}(m) + \frac{\tau_p^{2/p} + \|\beta\|^2}{n}\right).$$

The proof is completed by taking expectation on both sides of the last inequality.

We turn now on the proof of Inequality (2.8). Following the line of proof of Inequality (2.21), replacing $\widehat{\beta}_m^{(FPCR)}$ by $\widehat{\beta}_m^{(KB)}$, $\widehat{\Pi}_m$ by Π_m and $\widehat{\mathcal{M}_n}$ by $\mathcal{M}_n$, we obtain

$$\mathbb{E}\left[\|\widetilde{\beta}^{(KB)} - \beta\|_{\Gamma_n}^2 \mathbf{1}_{\overline{G}}\right] \leq C\left(\min_{m\in\mathcal{M}_n}\left\{\|\beta - \Pi_m\beta\|_{\Gamma_n}^2 + \mathrm{pen}(m)\right\} + \frac{\|\beta\|_\Gamma^2 + \tau_p^{2/p}}{n}\right),$$

where $C > 0$ depends only on θ, p and δ. Moreover, since $\widetilde{\beta}^{(KB)} = 0$ on the set $\overline{G}^{\complement}$,

$$\mathbb{E}\left[\|\widetilde{\beta}^{(KB)} - \beta\|_{\Gamma_n}^2 \mathbf{1}_{\overline{G}^{\complement}}\right] \leq \|\beta\|^2\mathbb{P}\left(\overline{G}^{\complement}\right).$$

The conclusion comes from Lemma 7. □

For sake of clarity, Lemma 2, which is the key of the previous result, is given below.

Lemma 2. *Suppose that Assumption **H1** is fulfilled. Let* $p(m,m') = 2(1+\delta)\frac{D_{m\vee m'}}{n}\sigma^2$, *then for all* $m \in \widehat{\mathcal{M}_n}$,

$$\sum_{m'\in\widehat{\mathcal{M}_n}} \mathbb{E}_{\mathbf{X}}\left[\left(\sup_{\substack{f\in\widehat{S}_{m\vee m'}\\ \|f\|_{\Gamma_n}=1}} \nu_n^2(f) - p(m,m')\right)_+\right] \leq \frac{C(p,\delta)}{n}\sigma^2.$$

The result also holds replacing $\widehat{\mathcal{M}_n}$ *by* $\mathcal{M}_n$ *and* $\widehat{S}_m$ *by* S_m.

Proof of Lemma 2. The proof of this lemma relies mainly on a result of Baraud (2000) given in Appendix B (Proposition 4, p.181). First denote by $\bar{f}_{\mathbf{X}} := (< f, X_1 >, ..., < f, X_n >)'$ and $\bar{\varepsilon} := (\varepsilon_1, ..., \varepsilon_n)'$. Remark that $\nu_n(f) = \bar{f}'_{\mathbf{X}}\bar{\varepsilon}/n$ and that, by usual properties of orthogonal projectors

$$\sup_{\substack{f\in\widehat{S}_m\\ \|f\|_{\Gamma_n}=1}} \nu_n(f) = \sup_{\substack{\alpha\in\widehat{s}_m\\ \alpha'\alpha=n}} \frac{\alpha'\bar{\varepsilon}}{n} = \frac{1}{\sqrt{n}} \sup_{\substack{\alpha\in\widehat{s}_m\\ \alpha'\alpha=1}} \alpha'\bar{\varepsilon} = \frac{1}{\sqrt{n}}(\Pi_{\widehat{s}_m}(\bar{\varepsilon})'\Pi_{\widehat{s}_m}\bar{\varepsilon})^{1/2} = \frac{1}{\sqrt{n}}\left(\bar{\varepsilon}'\Pi_{\widehat{s}_m}\bar{\varepsilon}\right)^{1/2},$$

where $\widehat{s}_m$ denotes the subspace of $\mathbb{R}^n$ defined by

$$\widehat{s}_m := \left\{\alpha \in \mathbb{R}^n,\ \exists f \in \widehat{S}_m,\ \alpha = \bar{f}_{\mathbf{X}}\right\}$$

and $\Pi_{\widehat{s}_m}$ is the orthogonal projector onto $\widehat{s}_m$.

Then, by Assumption **H1**, applying Proposition 4, p.181 in Appendix B with $\widetilde{A} = \Pi_{\widehat{s}_m}$ we obtain, for all $x > 0$,

$$\mathbb{P}_{\mathbf{X}}\left(n \sup_{\substack{f\in\widehat{S}_m\\ \|f\|_{\Gamma_n}=1}} \nu_n^2(f) \geq D_m\sigma^2 + 2\sigma^2\sqrt{D_m x} + \sigma^2 x\right) \leq C(p)\sigma^p\tau_p\frac{D_m}{x^{p/2}},$$

where $\mathbb{P}_{\mathbf{X}}$ stands for the probability given $\mathbf{X}$. Then for all $\delta > 0$ remark that $\sqrt{D_m x} \leq \delta D_m + \delta^{-1}x$ we obtain

$$\mathbb{P}_{\mathbf{X}}\left(\sup_{\substack{f\in\widehat{S}_m\\ \|f\|_{\Gamma_n}=1}} \nu_n^2(f) \geq (1+\delta)\frac{D_m\sigma^2}{n} + (1+\delta^{-1})\frac{\sigma^2 x}{n}\right) \leq C(p)\frac{D_m}{x^{p/2}}.$$

Set

$$Q_{m\vee m'} := \left(\sup_{\substack{f\in\widehat{S}_{m\vee m'}\\ \|f\|_{\Gamma_n}=1}} \nu_n^2(f) - p(m,m')\right)_+.$$

We have, for all $m, m' \in \mathcal{M}_n$

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}\left[Q_{m\vee m'}\right] &= \int_0^{+\infty} \mathbb{P}_{\mathbf{X}}\left(Q_{m\vee m'} \geq t\right) dt \\ &\leq C(p)\frac{\sigma^p}{n^{p/2}} \int_0^{+\infty} \frac{dt}{\left(t + \frac{\sigma^2 D_{m\vee m'}}{n}(1+\delta)\right)^{p/2}} \leq C'(p,\delta)\frac{\sigma^2}{n} D_{m\vee m'}^{1-p/2}. \qquad (2.23)\end{aligned}$$

Now remark that, since $(D_m)_{m\in\widehat{\mathcal{M}_n}}$ is strictly increasing and $D_1 \geq 1$, we have $\widehat{N}_n \leq D_{\widehat{N}_n}$,

$$\sum_{m\in\widehat{\mathcal{M}_n}} D_{m\vee m'}^{1-p/2} \leq \widehat{N}_n D_{\widehat{N}_n}^{1-p/2} \leq D_{\widehat{N}_n}^{2-p/2} \leq 1,$$

since $p > 4$. Hence Inequality (2.23) ends the proof. □

2.5.3 Proof of Theorem 1

Proof. The core of the proof relies on the bounds on the empirical risk given in Proposition 1. It remains to replace the empirical risk by the risk associated to the prediction error in order to obtain the final oracle-inequality. For this, we first introduce the following family of sets:

$$\Delta_m := \left\{\forall f \in S_m, \|f\|_\Gamma^2 \leq \rho_0 \|f\|_{\Gamma_n}^2\right\}; \qquad (2.24)$$

$$\widehat{\Delta}_m := \left\{\forall f \in \widehat{S}_m, \|f\|_\Gamma^2 \leq \rho_0 \|f\|_{\Gamma_n}^2\right\}. \qquad (2.25)$$

where $\rho_0 > 1$ is a constant. The following equalities hold:

$$\begin{aligned}\mathbb{E}[\|\widetilde{\beta}^{(KB)} - \beta\|_\Gamma^2] &= \mathbb{E}\left[\|\widetilde{\beta}^{(KB)} - \beta\|_\Gamma^2 \mathbf{1}_{\Delta_{N_n}}\right] + \mathbb{E}\left[\|\widetilde{\beta}^{(KB)} - \beta\|_\Gamma^2 \mathbf{1}_{\Delta_{N_n}^{\complement}}\right], \\ \mathbb{E}[\|\widetilde{\beta}^{(FPCR)} - \beta\|_\Gamma^2] &= \mathbb{E}\left[\|\widetilde{\beta}^{(FPCR)} - \beta\|_\Gamma^2 \mathbf{1}_{\left\{\widehat{\Delta}_{\widehat{N}_n^{(FPCR)}} \cap \widehat{\Delta}_{N_n}\right\}}\right] \\ &\quad + \mathbb{E}\left[\|\widetilde{\beta}^{(FPCR)} - \beta\|_\Gamma^2 \mathbf{1}_{\left\{\widehat{\Delta}_{\widehat{N}_n^{(FPCR)}}^{\complement} \cup \widehat{\Delta}_{N_n}^{\complement}\right\}}\right],\end{aligned}$$

where, for a set A, we denote by $A^{\complement}$ its complement.

Lemma 3 (for the estimator $\widetilde{\beta}^{(FPCR)}$) and Lemma 4 (for $\widetilde{\beta}^{(KB)}$) given below allows to bound the second terms of these equalities. Thus the end of the proof will be devoted to upper-bound the first terms.

We start with the proof for the upper-bound on $\mathbb{E}\left[\|\widetilde{\beta}^{(FPCR)} - \beta\|_\Gamma^2 \mathbf{1}_{\left\{\widehat{\Delta}_{\widehat{N}_n^{(FPCR)}} \cap \widehat{\Delta}_{N_n}\right\}}\right]$.

We first prove that, for all $m = 1, \ldots, N_n$,

$$\begin{aligned}\|\widetilde{\beta}^{(FPCR)} - \beta\|_\Gamma \mathbf{1}_{\left\{\widehat{\Delta}_{\widehat{N}_n^{(FPCR)}} \cap \widehat{\Delta}_{N_n}\right\}} &\leq \sqrt{\rho_0}\|\widetilde{\beta}^{(FPCR)} - \beta\|_{\Gamma_n} \\ &\quad + \sqrt{\rho_0}\|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n} + \|\beta - \widehat{\Pi}_m\beta\|_\Gamma. \qquad (2.26)\end{aligned}$$

Indeed, if $\widehat{N}_n^{(FPCR)} > N_n$, for all $m = 1, \ldots, \widehat{N}_n^{(FPCR)}$ (and hence for all $m = 1, \ldots, N_n$),

$$\begin{aligned}\|\widetilde{\beta}^{(FPCR)} - \beta\|_\Gamma \mathbf{1}_{\left\{\widehat{\Delta}_{\widehat{N}_n^{(FPCR)}} \cap \widehat{\Delta}_{N_n}\right\}} &\leq \|\widetilde{\beta}^{(FPCR)} - \widehat{\Pi}_m\beta\|_\Gamma \mathbf{1}_{\widehat{\Delta}_{\widehat{N}_n^{(FPCR)}}} \\ &\quad + \|\beta - \widehat{\Pi}_m\beta\|_\Gamma \\ &\leq \sqrt{\rho}_0\|\widetilde{\beta}^{(FPCR)} - \beta\|_{\Gamma_n} + \sqrt{\rho}_0\|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n} + \|\beta - \widehat{\Pi}_m\beta\|_\Gamma.\end{aligned}$$

where the last inequality comes from the fact that, for all $m \leq \widehat{N}_n^{(FPCR)}$, $\widehat{S}_m \subset \widehat{S}_{\widehat{N}_n^{(FPCR)}}$ and $\widetilde{\beta}^{(FPCR)} - \widehat{\Pi}_m\beta \in \widehat{S}_{\widehat{N}_n^{(FPCR)}}$ and the definition of $\widehat{\Delta}_{\widehat{N}_n^{(FPCR)}}$. Now, if $\widehat{N}_n^{(FPCR)} \leq N_n$, we have $\widehat{S}_{\widehat{N}_n^{(FPCR)}} \subset \widehat{S}_{N_n}$. Hence, for all $m = 1, \ldots, N_n$,

$$\|\widetilde{\beta}^{(FPCR)} - \beta\|_\Gamma \mathbf{1}_{\left\{\widehat{\Delta}_{\widehat{N}_n^{(FPCR)}} \cap \widehat{\Delta}_{N_n}\right\}} \leq \|\widetilde{\beta}^{(FPCR)} - \beta\|_\Gamma \mathbf{1}_{\widehat{\Delta}_{N_n}}$$

and the same reasoning than the one of the case $\widehat{N}_n^{(FPCR)} > N_n$ holds, remarking that here $\widetilde{\beta}^{(FPCR)} - \widehat{\Pi}_m\beta \in \widehat{S}_{N_n}$. Then Inequality (2.26) is true in both cases.

By Proposition 1, for all $m = 1, \ldots, N_n$:

$$\mathbb{E}\left[\|\widetilde{\beta}^{(FPCR)} - \beta\|_{\Gamma_n}^2\right] \leq C(p, \theta, \delta)\left(\mathbb{E}\left[\|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2\right] + \mathrm{pen}(m) + \frac{\tau_p^{2/p} + \|\beta\|^2}{n}\right)$$

and by Equation (2.26)

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\beta}^{(FPCR)} - \beta\|_\Gamma^2 \mathbf{1}_{\left\{\widehat{\Delta}_{\widehat{N}_n^{(FPCR)}} \cap \widehat{\Delta}_{N_n}\right\}}\right] &\leq C(p, \theta, \delta, \rho_0)\Big(\mathbb{E}[\|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2] \\ &\quad + \mathbb{E}[\|\beta - \widehat{\Pi}_m\beta\|_\Gamma^2] + \mathrm{pen}(m) + \frac{\tau_p^{2/p} + \|\beta\|^2}{n}\Big).\end{aligned}$$

With Lemma 19 p.170 we have (under **H1**, **H2** and **H4**) that

$$\mathbb{E}\left[\|\beta - \widehat{\Pi}_m\beta\|_{\Gamma_n}^2\right] \leq 4\mathbb{E}\left[\|\beta - \widehat{\Pi}_m\beta\|_\Gamma^2\right] + C\frac{D_m}{n}.$$

This implies, for all $m = 1, \ldots, N_n$,

$$\mathbb{E}\left[\|\widetilde{\beta}^{(FPCR)} - \beta\|_\Gamma^2 \mathbf{1}_{\left\{\widehat{\Delta}_{\widehat{N}_n^{(FPCR)}} \cap \widehat{\Delta}_{N_n}\right\}}\right] \leq C\left(\mathbb{E}[\|\beta - \widehat{\Pi}_m\beta\|_\Gamma^2] + \mathrm{pen}(m) + \frac{\tau_p^{2/p} + \|\beta\|^2}{n}\right),$$

with $C > 0$ independent of β and n.

The upper-bound on $\mathbb{E}\left[\|\widetilde{\beta}^{(KB)} - \beta\|_\Gamma^2 \mathbf{1}_{\Delta_{N_n}}\right]$ comes from similar arguments. The following inequality, verified for all $m = 1, \ldots, N_n$,

$$\|\widetilde{\beta}^{(KB)} - \beta\|_\Gamma \mathbf{1}_{\Delta_{N_n}} \leq \sqrt{\rho}_0\|\widetilde{\beta}^{(KB)} - \beta\|_{\Gamma_n} + \sqrt{\rho}_0\|\beta - \Pi_m\beta\|_{\Gamma_n} + \|\beta - \Pi_m\beta\|_\Gamma,$$

is obtained as Inequality (2.26) for $\widetilde{\beta}^{(FPCR)}$. Now Proposition 1 implies that

$$\mathbb{E}\left[\|\widetilde{\beta}^{(KB)}-\beta\|^2_{\Gamma_n}\right] \leq C(p,\theta,\delta)\left(\mathbb{E}\left[\|\beta-\Pi_m\beta\|^2_{\Gamma_n}\right]+\mathrm{pen}(m)+\frac{\tau_p^{2/p}+\|\beta\|^2}{n}\right).$$

Moreover, since

$$\mathbb{E}\left[\|\beta-\Pi_m\beta\|^2_{\Gamma_n}\right]=\|\beta-\Pi_m\beta\|^2_{\Gamma},$$

we have

$$\mathbb{E}\left[\|\widetilde{\beta}^{(KB)}-\beta\|_{\Gamma}\mathbf{1}_{\Delta_{N_n}}\right] \leq C\left(\|\beta-\Pi_m\beta\|^2_{\Gamma}+\mathrm{pen}(m)+\frac{\tau_p^{2/p}+\|\beta\|^2}{n}\right),$$

with $C>0$ independent of β and n.

Finally, with Lemma 3 and Lemma 4 below, we obtain inequalities (2.10) and (2.11) of Theorem 1. □

Lemma 3. *For all $\beta\in\mathbb{H}$, under assumptions **H2**, **H3** and **H4** and if the decreasing rate of the sequence $(\lambda_j)_{j\geq 1}$ is given by **(P)** or **(E)** we have:*

$$\mathbb{E}\left[\|\widetilde{\beta}^{(FPCR)}-\beta\|_{\Gamma}\mathbf{1}_{\left\{\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}}\cup\widehat{\Delta}^{\complement}_{N_n}\right\}}\right] \leq \frac{C}{n}(1+\|\beta\|^2_{\Gamma}),$$

with $C>0$ independent of β and n.

Proof. First recall that with Equation (2.6) and as $Y_i=<\beta,X_i>+\varepsilon_i$, for all $m\in\widehat{\mathcal{M}_n}$

$$\widehat{\beta}_m^{(FPCR)}=\sum_{j=1}^{D_m}\frac{1}{n}\sum_{i=1}^{n}Y_i\frac{<X_i,\widehat{\psi}_j>}{\widehat{\lambda}_j}\widehat{\psi}_j=\widehat{\Pi}_m\beta+R_m,$$

where

$$R_m:=\sum_{j=1}^{D_m}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\frac{<X_i,\widehat{\psi}_j>}{\widehat{\lambda}_j}\widehat{\psi}_j.$$

Then

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\beta}^{(FPCR)}-\beta\|^2_{\Gamma}\mathbf{1}_{\left\{\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}}\cup\widehat{\Delta}^{\complement}_{N_n}\right\}}\right] &\leq 2\mathbb{E}\left[\|\widehat{\Pi}_{\widehat{m}^{(FPCR)}}\beta-\beta\|^2_{\Gamma}\mathbf{1}_{\left\{\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}}\cup\widehat{\Delta}^{\complement}_{N_n}\right\}}\right]\\ &\quad+2\mathbb{E}\left[\|R_{\widehat{m}^{(FPCR)}}\|^2_{\Gamma}\mathbf{1}_{\left\{\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}}\cup\widehat{\Delta}^{\complement}_{N_n}\right\}}\right]\\ &\leq 2\|\beta\|^2_{\Gamma}\mathbb{P}\left(\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}}\cup\widehat{\Delta}^{\complement}_{N_n}\right)+2\mathbb{E}\left[\|R_{\widehat{m}^{(FPCR)}}\|^2_{\Gamma}\mathbf{1}_{\left\{\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}}\cup\widehat{\Delta}^{\complement}_{N_n}\right\}}\right].\end{aligned}$$

The first term can be easily bounded using the results of Lemma 20 in Appendix A, p.172. Then we focus on the second term, the idea is to bound the quantity $\|R_{\widehat{m}^{(FPCR)}}\|_{\Gamma}$ by

$\|R_{\widehat{m}^{(FPCR)}}\|$ which can be written simply,

$$\begin{aligned}\mathbb{E}\left[\|R_{\widehat{m}^{(FPCR)}}\|_\Gamma^2 \mathbf{1}_{\left\{\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \cup \widehat{\Delta}^{\complement}_{N_n}\right\}}\right] &\leq \rho(\Gamma)\mathbb{E}\left[\|R_{\widehat{m}^{(FPCR)}}\|^2 \mathbf{1}_{\left\{\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \cup \widehat{\Delta}^{\complement}_{N_n}\right\}}\right] \\ &= \rho(\Gamma)\mathbb{E}\left[\sum_{j=1}^{D_{\widehat{m}^{(FPCR)}}} < R_{\widehat{m}^{(FPCR)}}, \widehat{\psi}_j >^2 \mathbf{1}_{\left\{\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \cup \widehat{\Delta}^{\complement}_{N_n}\right\}}\right].\end{aligned}$$

Now remark that, since $\widehat{\lambda}_j \geq \mathfrak{s}_n$, for all $j \leq D_{\widehat{m}^{(FPCR)}}$ (as $D_{\widehat{m}^{(FPCR)}} \leq D_{\widehat{N}_n^{(FPCR)}}$) and $D_{\widehat{m}^{(FPCR)}} \leq D_{\widehat{N}_n^{(FPCR)}} \leq 20\sqrt{n}$,

$$\begin{aligned}\sum_{j=1}^{D_{\widehat{m}^{(FPCR)}}} < R_{\widehat{m}^{(FPCR)}}, \widehat{\psi}_j >^2 &\leq \sum_{j=1}^{D_{\widehat{m}^{(FPCR)}}} \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \frac{< X_i, \widehat{\psi}_j >}{\widehat{\lambda}_j}\right)^2 \\ &\leq \mathfrak{s}_n^{-1} \sum_{j=1}^{D_{\widehat{m}^{(FPCR)}}} \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \frac{< X_i, \widehat{\psi}_j >}{\sqrt{\widehat{\lambda}_j}}\right)^2 \\ &\leq \mathfrak{s}_n^{-1} \sum_{j=1}^{20\lfloor\sqrt{n}\rfloor} \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \frac{< X_i, \widehat{\psi}_j >}{\sqrt{\widehat{\lambda}_j}}\right)^2 \mathbf{1}_{\{\widehat{\lambda}_j > 0\}}.\end{aligned}$$

Then we have, by independence of ε_i with X_i and $\widehat{\Delta}_{\widehat{N}_n^{(FPCR)}}$,

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}\left[\sum_{j=1}^{D_{\widehat{m}^{(FPCR)}}} < R_{\widehat{m}^{(FPCR)}}, \widehat{\psi}_j >^2\right] &\leq \frac{\sigma^2}{n^2}\mathfrak{s}_n^{-1} \sum_{j=1}^{20\lfloor\sqrt{n}\rfloor} \sum_{i=1}^n \frac{< X_i, \widehat{\psi}_j >^2}{\widehat{\lambda}_j} \mathbf{1}_{\{\widehat{\lambda}_j > 0\}} \\ &= \frac{\sigma^2}{n}\mathfrak{s}_n^{-1} \sum_{j=1}^{20\lfloor\sqrt{n}\rfloor} \frac{< \Gamma_n \widehat{\psi}_j, \widehat{\psi}_j >}{\widehat{\lambda}_j} \mathbf{1}_{\{\widehat{\lambda}_j > 0\}} \leq \frac{\sigma^2}{n}\mathfrak{s}_n^{-1} 20\sqrt{n},\end{aligned}$$

then

$$\mathbb{E}\left[\|R_{\widehat{m}^{(FPCR)}}\|_\Gamma^2 \mathbf{1}_{\left\{\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \cup \widehat{\Delta}^{\complement}_{N_n}\right\}}\right] \leq 20\rho(\Gamma)\frac{\sigma^2}{\sqrt{n}}\mathfrak{s}_n^{-1}\mathbb{P}\left(\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \cup \widehat{\Delta}^{\complement}_{N_n}\right).$$

Now from Lemma 20 in Appendix A (p.172) we have

$$\mathbb{P}\left(\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \cup \widehat{\Delta}^{\complement}_{N_n}\right) \leq C/n^6,$$

where $C > 0$ is independent of β and n and by definition of $\mathfrak{s}_n = (2/n^2)(1 - 1/\ln^2 n)$ we get

$$\mathbb{E}\left[\|R_{\widehat{m}^{(FPCR)}}\|_\Gamma^2 \mathbf{1}_{\left\{\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \cup \widehat{\Delta}^{\complement}_{N_n}\right\}}\right] \leq C' n^{-9/2},$$

where C' is a constant.

Finally

$$\mathbb{E}\left[\|\widetilde{\beta}^{(FPCR)}-\beta\|_\Gamma \mathbf{1}_{\left\{\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}}\cup\widehat{\Delta}^{\complement}_{N_n}\right\}}\right] \leq C^{(3)}\left(\|\beta\|_\Gamma^2 n^{-6}+n^{-9/2}\right),$$

where $C^{(3)}>0$ does not depend on β nor on n. □

Lemma 4. *Under Assumption **H2**, if $\mathbb{E}[\langle\beta,X_1\rangle^4]<+\infty$, there exists a constant C' depending only on ρ_0, K, c and v such that:*

$$\mathbb{E}[\|\widetilde{\beta}^{(KB)}-\beta\|_\Gamma^2 \mathbf{1}_{\Delta^{\complement}_{N_n}}] \leq \frac{C'}{n}(\mathbb{E}[<\beta,X_1>^4]^{1/2}+\|\beta\|_\Gamma^2+1).$$

Proof. First, by triangular inequality,

$$\begin{aligned}\mathbb{E}[\|\widetilde{\beta}^{(KB)}-\beta\|_\Gamma^2 \mathbf{1}_{\Delta^{\complement}_{N_n}}] &\leq 2\mathbb{E}\left[\left(\left\|\widetilde{\beta}^{(KB)}\right\|_\Gamma^2+\|\beta\|_\Gamma^2\right)\mathbf{1}_{\Delta^{\complement}_{N_n}}\right]\\ &= 2\mathbb{E}\left[\left\|\widehat{\beta}^{(KB)}_{\widehat{m}}\right\|_\Gamma^2 \mathbf{1}_{\Delta^{\complement}_{N_n}}\right]+2\|\beta\|_\Gamma^2\mathbb{P}\left(\Delta^{\complement}_{N_n}\right). \qquad (2.27)\end{aligned}$$

First, for any function $f=\sum_{j=1}^{D_{N_n}}\alpha_j\psi_j\in\mathcal{S}_n\backslash\{0\}$, if $\overline{G}$ is verified

$$\|f\|_{\Gamma_n}^2=\alpha'\widehat{\Phi}_{N_n}\alpha\geq\widetilde{\lambda}_{N_n}\alpha'\alpha\geq\mathfrak{s}_n\alpha'\alpha/2,$$

where we recall that $\widehat{\Phi}_{N_n}=(\langle\Gamma_n\psi_j,\psi_k\rangle)_{1\leq j,k\leq D_{N_n}}$ and that $\widetilde{\lambda}_{N_n}$ is the smallest eigenvalue of $\widehat{\Phi}_{N_n}$. Moreover,

$$\|f\|_\Gamma^2=\sum_{j=1}^{D_{N_n}}\lambda_j\alpha_j^2\leq\rho(\Gamma)\alpha'\alpha.$$

Then $\|f\|_\Gamma^2\leq\frac{2\rho(\Gamma)}{\mathfrak{s}_n}\|f\|_{\Gamma_n}^2$. Now, taking $f=\widetilde{\beta}^{(KB)}$ and remarking that, since $\widetilde{\beta}^{(KB)}=0$ on $\overline{G}^{\complement}$,

$$\mathbb{E}\left[\left\|\widehat{\beta}^{(KB)}_{\widehat{m}}\right\|_\Gamma^2 \mathbf{1}_{\Delta^{\complement}_{N_n}}\right]=\mathbb{E}\left[\left\|\widehat{\beta}^{(KB)}_{\widehat{m}}\right\|_\Gamma^2 \mathbf{1}_{\Delta^{\complement}_{N_n}\cap\overline{G}}\right],$$

we obtain:

$$\mathbb{E}\left[\|\widetilde{\beta}^{(KB)}\|_\Gamma^2\mathbf{1}_{\Delta^{\complement}_{N_n}}\right] \leq \frac{2\rho(\Gamma)}{\mathfrak{s}_n}\mathbb{E}\left[\|\widetilde{\beta}^{(KB)}\|_{\Gamma_n}^2\mathbf{1}_{\Delta^{\complement}_{N_n}}\right]. \qquad (2.28)$$

Now, since $\widehat{\beta}^{(KB)}_m$ is a mean-square-type estimator, the vector $(\langle\widehat{\beta}^{(KB)}_m,X_1\rangle,\cdots,\langle\widehat{\beta}^{(KB)}_m,X_n\rangle)'$ can be seen as the orthogonal projection (w.r.t the Euclidean scalar product on $\mathbf{R}^n$) of the vector $(Y_1,\cdots,Y_n)'$ on the subspace $\{(\langle f,X_1\rangle,\cdots,\langle f,X_n\rangle)',f\in S_m\}$. Since the squared empirical norm $\|\cdot\|_n^2$ corresponds to the Euclidean norm up to the multiplicative factor $1/n$,

we deduce that:

$$\sum_{i=1}^{n}\langle\widehat{\beta}_m^{(KB)},X_i\rangle^2 = n\|\widehat{\beta}_m^{(KB)}\|_{\Gamma_n}^2 \leq \sum_{i=1}^{n} Y_i^2, \text{ for all } m,$$

as the norm of the vector $(Y_1,\cdots,Y_n)'$ is larger than the norm of its projection. Hence

$$\sum_{i=1}^{n}\langle\widetilde{\beta}^{(KB)},X_i\rangle^2 = n\|\widetilde{\beta}^{(KB)}\|_{\Gamma_n}^2 \leq \sum_{i=1}^{n} Y_i^2.$$

Then, we can use that $Y_i = \langle\beta,X_i\rangle + \varepsilon_i$ and have:

$$\|\widetilde{\beta}^{(KB)}\|_{\Gamma_n}^2 \leq 2\|\beta\|_{\Gamma_n}^2 + \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i^2.$$

By gathering the last inequality and inequalities (2.27) and (2.28), we obtain:

$$\begin{aligned}\mathbb{E}[\|\widetilde{\beta}^{(KB)}-\beta\|_\Gamma^2\mathbf{1}_{\Delta_{N_n}^{\complement}}] &\leq \frac{8\rho(\Gamma)}{\mathfrak{s}_n}\mathbb{E}\left[\left(\|\beta\|_{\Gamma_n}^2+\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right)\mathbf{1}_{\Delta_{N_n}^{\complement}}\right]\\ &\quad +2\|\beta\|_\Gamma^2\mathbb{P}\left(\Delta_{N_n}^{\complement}\right).\end{aligned}$$

The ε_i's are independent of the X_i's, and the set $\Delta_{\widehat{N}_n^{(KB)}}^{\complement}$ depends only on the X_i's so that:

$$\mathbb{E}\left[\frac{1}{n}\left(\sum_{i=1}^{n}\varepsilon_i^2\right)\mathbf{1}_{\Delta_{N_n}^{\complement}}\right] = \sigma^2\mathbb{P}\left(\Delta_{N_n}^{\complement}\right)$$

On the other hand, by Cauchy-Schwarz Inequality, we have:

$$\begin{aligned}\mathbb{E}\left[\|\beta\|_{\Gamma_n}^2\mathbf{1}_{\Delta_{N_n}^{\complement}}\right] &\leq \mathbb{E}[\|\beta\|_{\Gamma_n}^4]^{1/2}\sqrt{\mathbb{P}\left(\Delta_{N_n}^{\complement}\right)}\\ &= \left(\frac{1}{n}\mathbb{E}[\langle\beta,X_1\rangle^4]+\frac{n-1}{n}\|\beta\|_\Gamma^4\right)^{1/2}\sqrt{\mathbb{P}\left(\Delta_{N_n}^{\complement}\right)}\\ &\leq \left(\frac{1}{\sqrt{n}}[E\langle\beta,X_1\rangle^4]^{1/2}+\|\beta\|_\Gamma^2\right)\sqrt{\mathbb{P}\left(\Delta_{N_n}^{\complement}\right)}\end{aligned}$$

As $\mathbb{P}\left(\Delta_{N_n}^{\complement}\right) \leq \sqrt{\mathbb{P}\left(\Delta_{N_n}^{\complement}\right)}$ and $\mathfrak{s}_n \leq 2$, we get:

$$\begin{aligned}\mathbb{E}[\|\widetilde{\beta}^{(KB)}-\beta\|_\Gamma^2\mathbf{1}_{\Delta_{N_n}^{\complement}}] &\leq \frac{8\sqrt{\mathbb{P}\left(\Delta_{N_n}^{\complement}\right)}}{\mathfrak{s}_n}\Big(\rho(\Gamma)[\mathbb{E}\langle\beta,X_1\rangle^4]^{1/2}/\sqrt{n}+\\ &\quad \rho(\Gamma)\|\beta\|_\Gamma^2+\rho(\Gamma)\sigma^2+\|\beta\|_\Gamma^2\Big).\end{aligned}$$

Lemma 5 below ends the proof. □

Lemma 5. *Under Assumption **H2**, for all $\rho_0 > 1$,*

$$\mathbb{P}\left(\Delta_{N_n}^{\complement}\right) \leq \frac{C}{n^4},$$

where $C > 0$ depends only on ρ_0 and b.

Proof. By definition of Δ_{N_n}:

$$\Delta_{N_n}^{\complement} = \left\{ \inf_{f \in S_{N_n} \setminus \{0\}} \frac{\|f\|_{\Gamma_n}^2}{\|f\|_{\Gamma}^2} < \rho_0^{-1} \right\}.$$

Now remark that

$$\begin{aligned} \inf_{f \in S_{N_n} \setminus \{0\}} \frac{\|f\|_{\Gamma_n}^2}{\|f\|_{\Gamma}^2} &= \inf_{\alpha \in \mathbb{R}^{D_{N_n}} \setminus \{0\}} \frac{\alpha' \widehat{\Phi}_{N_n} \alpha}{\alpha' \Lambda_{N_n} \alpha} = \inf_{\alpha \in \mathbb{R}^{D_{N_n}} \setminus \{0\}} \frac{\alpha' \Lambda_{N_n}^{-1/2} \widehat{\Phi}_{N_n} \Lambda_{N_n}^{-1/2} \alpha}{\alpha' \alpha} \\ &= \rho\left(\Lambda_{N_n}^{1/2} \widehat{\Phi}_{N_n}^{-1} \Lambda_{N_n}^{1/2}\right)^{-1} = \rho(\Psi_{N_n}^{-1})^{-1}, \end{aligned}$$

where Λ_{N_n} is the diagonal matrix with entries $(\lambda_1, \ldots, \lambda_{D_{N_n}})$, we recall that $\widehat{\Phi}_{N_n} = (\langle \Gamma_n \psi_j, \psi_k \rangle)_{1 \leq j,k \leq D_{N_n}}$ and we denote by

$$\Psi_{N_n} := \Lambda_{N_n}^{-1/2} \widehat{\Phi}_{N_n} \Lambda_{N_n}^{-1/2} = \left(\frac{\langle \Gamma_n \psi_j, \psi_k \rangle}{\sqrt{\lambda_j \lambda_k}} \right)_{1 \leq j,k \leq D_{N_n}}.$$

Now remark that $\rho(\Psi_{N_n}^{-1})^{-1}$ is the smallest eigenvalue of Ψ_{N_n} ; this implies that

$$1 - \rho(\Psi_{N_n}^{-1})^{-1} \leq \rho\left(I - \Psi_{N_n}\right),$$

then applying Lemma 6 given below, with $m = N_n$ and $t = 1 - \rho_0^{-1}$, we get

$$\begin{aligned} \mathbb{P}\left(\Delta_{N_n}^{\complement}\right) &= \mathbb{P}\left(1 - \rho\left(\Psi_{N_n}^{-1}\right)^{-1} > 1 - \rho_0^{-1}\right) \leq \mathbb{P}\left(\rho\left(I - \Psi_{N_n}\right) > 1 - \rho_0^{-1}\right) \\ &\leq 2 D_{N_n}^2 \exp\left(- \frac{n}{D_{N_n}^2} \frac{(1 - \rho_0^{-1})^2}{3b} \right), \end{aligned}$$

and the conclusion follows from the fact that $D_{N_n} \leq 20\sqrt{n / \ln^3 n}$. □

Lemma 6. *Define, for $m \in \mathbb{N}^*$, the matrix*

$$\Psi_m := \left(\frac{\langle \Gamma_n \psi_j, \psi_k \rangle}{\sqrt{\lambda_j \lambda_k}} \right)_{1 \leq j,k \leq D_m}. \tag{2.29}$$

*If Assumption **H2** is verified, for all $t > 0$,*

$$\mathbb{P}(\rho(\Psi_m - I) > t) \leq 2 D_m^2 \exp\left(- \frac{n}{D_m^2} \frac{t^2}{3b} \right). \tag{2.30}$$

Proof. Define, for $j,k = 1, ..., D_m$:

$$Z_i^{(j,k)} = \frac{< \psi_j, X_i >}{\sqrt{\lambda_j}} \frac{< \psi_k, X_i >}{\sqrt{\lambda_k}},$$

we have, for all, j, k, $\mathbb{E}\left[Z_i^{(j,k)}\right] = \delta_{j,k}$ and then

$$\Psi_m - I = \left(\frac{1}{n} \sum_{i=1}^{n} Z_i^{(j,k)} - \mathbb{E}\left[Z_i^{(j,k)}\right] \right)_{1 \leq j,k \leq D_m}.$$

The aim is then to use Bernstein's Inequality to control $\rho(\Psi_m - I)$.

First remark that the trace of a matrix is equal to the sum of its eigenvalues (counted according to their algebraic multiplicities), then:

$$\rho\left(I - \Psi_m\right)^2 \leq \operatorname{tr}\left(\left(\Psi_m - I\right)^2\right) = \operatorname{tr}\left(\left(\Psi_m - I\right)' \left(\Psi_m - I\right)\right), \tag{2.31}$$

as $\Psi_m - I$ is symmetric. The last term of the inequality is equal to the sum of the squared coefficients of $\Psi_m - I$.
Hence, by (2.31) :

$$\rho(\Psi_m - I)^2 \leq \sum_{1 \leq j,k \leq D_m} \left(\frac{1}{n} \sum_{i=1}^{n} Z_i^{(j,k)} - \mathbb{E}\left[Z_i^{(j,k)}\right] \right)^2.$$

This gives:

$$\begin{aligned}
\mathbb{P}\left(\rho(\Psi_m - I) > t\right) &\leq \mathbb{P}\left(\sum_{1 \leq j,k \leq D_m} \left(\frac{1}{n} \sum_{i=1}^{n} Z_i^{(j,k)} - \mathbb{E}\left[Z_i^{(j,k)}\right] \right)^2 > t^2 \right) \\
&\leq \mathbb{P}\left(\bigcup_{1 \leq i,j \leq D_m} \left\{ \left(\frac{1}{n} \sum_{i=1}^{n} Z_i^{(j,k)} - \mathbb{E}\left[Z_i^{(j,k)}\right] \right)^2 > \frac{t^2}{D_m^2} \right\} \right) \\
&\leq \sum_{1 \leq j,k \leq D_m} \mathbb{P}\left(\left(\frac{1}{n} \sum_{i=1}^{n} Z_i^{(j,k)} - \mathbb{E}\left[Z_i^{(j,k)}\right] \right)^2 > \frac{t^2}{D_m^2} \right) \\
&\leq \sum_{1 \leq j,k \leq D_m} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} Z_i^{(j,k)} - \mathbb{E}\left[Z_i^{(j,k)}\right] \right| > \frac{t}{D_m} \right).
\end{aligned} \tag{2.32}$$

By Cauchy–Schwarz inequality and Assumption **H2**,

$$\operatorname{Var}\left(Z_i^{(j,k)}\right) \leq \mathbb{E}\left[Z_i^{(j,k)2}\right] \leq \mathbb{E}\left[\frac{\langle X_i, \psi_j \rangle^4}{\lambda_j^2}\right]^{1/2} \mathbb{E}\left[\frac{\langle X_i, \psi_k \rangle^4}{\lambda_k^2}\right]^{1/2} \leq 2b$$

and, for all $\ell \geq 2$,

$$\mathbb{E}\left[\left|Z_i^{(j,k)}\right|^\ell\right] \leq \mathbb{E}\left[\frac{\langle X_i, \psi_j\rangle^{2\ell}}{\lambda_j^\ell}\right]^{1/2} \mathbb{E}\left[\frac{\langle X_i, \psi_k\rangle^{2\ell}}{\lambda_k^\ell}\right]^{1/2} \leq \ell! b^{\ell-1} = \frac{\ell!}{2} v^2 b^{\ell-2}.$$

Then, we can apply Bernstein's Inequality with $v^2 = 2b$ and $b_0 = b$ (see Lemma 22 in Appendix B p.180) to the sequence $Z_1^{(j,k)}, ..., Z_n^{(j,k)}$, for all $j, k = 1, ..., D_m$. We obtain, since $D_m \geq 1$,

$$\mathbb{P}\left(\rho(\Psi_m - I) > t\right) \leq 2D_m^2 \exp\left(-\frac{n\left(\frac{t}{D_m}\right)^2}{2b + b\frac{t}{D_m}}\right) \leq 2D_m^2 \exp\left(-\frac{n}{D_m^2}\frac{t^2}{3b}\right).$$

This concludes the proof. □

Lemma 7. *If Assumption **H2** is fulfilled, then*

$$\mathbb{P}\left(\overline{G}^{\complement}\right) \leq C/n^4,$$

where $C > 0$ only depends on b.

Proof. We have, by definition of $\overline{G}$

$$\mathbb{P}\left(\overline{G}^{\complement}\right) \leq \sum_{m \in \mathcal{M}_n} \mathbb{P}\left(2\widetilde{\lambda}_m < \mathfrak{s}_n\right). \tag{2.33}$$

Let Ψ_m the matrix defined in Equation (2.29), p. 68, we have

$$\Psi_m = \Lambda_m^{-1/2} \widehat{\Phi}_m \Lambda_m^{-1/2}.$$

Then, let $\widetilde{\mu}_m$ be the smallest eigenvalue of Ψ_m, we prove in a first step that

$$\widetilde{\mu}_m \leq \lambda_{D_m}^{-1} \widetilde{\lambda}_m. \tag{2.34}$$

If $\widetilde{\mu}_m = 0$, then (2.34) is true since $\widetilde{\lambda}_m \geq 0$ (the matrix $\widehat{\Phi}_m$ is positive). Otherwise, we have $\widetilde{\mu}_m > 0$ and

$$\widetilde{\mu}_m = \rho\left(\Psi_m^{-1}\right)^{-1},$$

since all the eigenvalues of Ψ_m are non negative (for a matrix A, we denote $\rho(A) = \max\{|\lambda|,\ \lambda \text{ eigenvalue of } A\}$). Since Φ_m^{-1} is symmetric and positive, we have

$$\widetilde{\lambda}_m^{-1} = \rho\left(\Phi_m^{-1}\right) = |||\Phi_m^{-1}||| = |||\Lambda_m^{-1/2} \widehat{\Psi}_m^{-1} \Lambda_m^{-1/2}||| \leq |||\Lambda_m^{-1/2}|||^2\ |||\widehat{\Psi}_m^{-1}||| = \lambda_{D_m}^{-1} \widetilde{\mu}_m^{-1},$$

where, for a D_m by D_m squared matrix A, $|||A||| = \sup_{x \in \mathbb{R}^{D_m}} \{xAx'/xx'\}$. This implies (2.34).

From (2.34) and (2.33), we deduce that

$$\mathbb{P}\left(\overline{G}^{\complement}\right) \leq \sum_{m\in\mathcal{M}_n} \mathbb{P}\left(\widetilde{\mu}_m < \lambda_{D_m}^{-1}\mathfrak{s}_n/2\right) \leq \sum_{m\in\mathcal{M}_n} \mathbb{P}\left(\widetilde{\mu}_m < \lambda_{D_m}^{-1}\mathfrak{s}_n/2\right).$$

Now, remark that, for all $m \in \mathcal{M}_n$, since $\lambda_{D_m}^{-1}\mathfrak{s}_n/2 \leq n^{-2}\mathfrak{s}_n/2 = 1 - 1/\sqrt{\ln(n)} < 1$,

$$\begin{aligned}\widetilde{\mu}_m < \lambda_{D_m}^{-1}\mathfrak{s}_n/2 &\Rightarrow 1-\widetilde{\mu}_m > 1-\lambda_{D_m}^{-1}\mathfrak{s}_n/2\\ &\Rightarrow |1-\widetilde{\mu}_m| > |1-\lambda_{D_m}^{-1}\mathfrak{s}_n/2| \Rightarrow \rho(1-\Psi_m) > |1-\lambda_{D_m}^{-1}\mathfrak{s}_n/2| = 1/\sqrt{\ln(n)}.\end{aligned}$$

Hence,

$$\mathbb{P}\left(\overline{G}^{\complement}\right) \leq \sum_{m\in\mathcal{M}_n} \mathbb{P}\left(\rho(1-\Psi_m) > \sqrt{\ln n}\right).$$

And applying Lemma 6, we have, since, for all $m \in \mathcal{M}_n$, $D_m \leq D_{N_n} \leq 20\sqrt{n/(\ln(n))^3}$

$$\mathbb{P}\left(\overline{G}^{\complement}\right) \leq \sum_{m\in\mathcal{M}_n} 2D_m^2 \exp\left(-\frac{n}{3b\ln(n)D_m^2}\right) \leq \frac{16000n^{3/2}}{(\ln(n))^{9/2}} \exp\left(-\frac{(\ln(n))^2}{1200b}\right).$$

□

2.5.4 Proof of Theorem 2

Let us start with the polynomial case **(P)**. By Theorem 1, we have, for all $\beta \in \mathcal{W}_R^r$,

$$\mathbb{E}[\|\widetilde{\beta}^{(FPCR)} - \beta\|_\Gamma^2] \leq C_1\left(\min_{m\in\mathcal{M}_n}\left(\mathbb{E}[\|\beta - \widehat{\Pi}_m\beta\|_\Gamma^2] + \mathrm{pen}(m)\right) + \frac{1}{n}\right),$$

and

$$\mathbb{E}[\|\widetilde{\beta}^{(KB)} - \beta\|_\Gamma^2] \leq C_2\left(\min_{m\in\mathcal{M}_n}\left(\mathbb{E}[\|\beta - \Pi_m\beta\|_\Gamma^2] + \mathrm{pen}(m)\right) + \frac{1}{n}\right),$$

with C_1, C_2 independent of β and n. First remark that since $\beta \in \mathcal{W}_r^R$,

$$\mathbb{E}[\|\beta - \Pi_m\beta\|_\Gamma^2] = \sum_{j\geq 1}\lambda_j < \beta, \psi_j >^2 \leq D_m^{-a-r}\sum_{j\geq D_m} j^r < \beta, \psi_j >^2 \leq R^2 D_m^{-a-r}.$$

We can see that it is possible to define a sequence of integers $(m_n^*)_{n\in\mathbb{N}^*}$ such that

$$cn^{1/a+r+1} \leq D_{m_n^*} \leq Cn^{1/a+r+1} \text{ and } D_{m_n^*} \leq N_n \text{ for all } n \in \mathbb{N}^*, \tag{2.35}$$

with c, C two positive constants. Now, considerations above lead us to

$$\sup_{\beta\in\mathcal{W}_r^R} \mathbb{E}[\|\beta - \Pi_{m_n^*}\beta\|_\Gamma^2] \leq C' n^{-\frac{a+r}{a+r+1}},$$

and

$$\mathrm{pen}(m_n^*) \leq C'' n^{-\frac{a+r}{a+r+1}}.$$

This leads to the expected bound:

$$\mathbb{E}[\|\widetilde{\beta}^{(KB)} - \beta\|_\Gamma^2] \leq C^{(3)} n^{-\frac{a+r}{a+r+1}},$$

with C', C'' and $C^{(3)}$ some positive constants.

We deal now with $\widetilde{\beta}^{(FPCR)}$, the idea is the same but the bias is more complex. By Lemma 18 p.165:

$$\begin{aligned}\mathbb{E}\left[\|\beta - \widehat{\Pi}_m\beta\|_\Gamma^2\right] &\leq 2\|\beta - \Pi_m\beta\|_\Gamma^2 + C_1\frac{\ln^3(D_m+1)}{n} D_m^{\max\{(1-r)_+, 2-a-r\}} \\ &\quad + C_2\frac{\ln^9 n}{n^2} D_m^{(4+(a-r)_+-2a)_++2},\end{aligned}$$

with $C_1, C_2 > 0$, independent of β and n and, for all $\alpha \in \mathbb{R}$, $(\alpha)_+ = \max\{\alpha, 0\}$. Then taking $m = m_n^*$ as in Equation (2.35) allows to obtain the same rate provided that $a + r/2 > 2$.

The exponential case **(E)** is treated similarly taking $c\ln^{1/a} n \leq D_{m_n^*} \leq C\ln^{1/a} n$.

□

Chapitre 3

Estimateur à noyau adaptatif de la fonction de répartition conditionnelle

En collaboration avec Gaëlle Chagny, LMRS, Université de Rouen.

L'objectif de ce chapitre est d'étudier, d'un point de vue non-asymptotique, l'estimateur de type Nadaraya-Watson de la fonction de répartition conditionnelle de Y sachant X proposé pour la première fois par Ferraty, Laksaci et Vieu (2006) et de proposer une procédure de sélection de la fenêtre. Nous étudions dans un premier temps le biais et la variance de l'estimateur pour un risque ponctuel et un risque intégré. Nous proposons ensuite un critère de sélection de la fenêtre dans l'esprit de la méthode de Goldenshluger et Lepski (2011). L'estimateur obtenu est adaptatif et atteint la vitesse de convergence minimax : nous établissons des bornes non-asymptotiques et calculons les vitesses de convergence sous plusieurs hypothèses portant sur le comportement de la probabilité de petite boule au voisinage de 0. Nous montrons également des bornes inférieures sur la vitesse de convergence pour les deux risques (ponctuel et intégré). Nous discutons ensuite du choix de la norme ou de la semi-norme dans le noyau en reliant cette question à la projection des données dans un espace de dimension finie. Le critère de sélection de la fenêtre est finalement étudié sur des données simulées et réelles.

Ce chapitre est une version modifiée de l'article :
Chagny, G. et Roche, A. (2014). Adaptive and minimax estimation of the cumulative distribution function given a functional covariate, HAL : hal-00931228, en révision.

Sommaire

3.1 Introduction

We are interested in explaining the relationship between a random variable $X \in \mathbb{H}$ and a scalar quantity Y. The link between the predictor X and the response Y is classically described by regression analysis. However, this can also be achieved by estimating the conditional distribution function of the variable Y which permits to visualize the entire conditional distribution of Y given X (whereas the regression function gives only the conditional expectation). The target function we want to recover is the conditional cumulative distribution function (conditional c.d.f. in the sequel) of Y given X defined by

$$F^x(y) := \mathbb{P}(Y \leq y | X = x), \quad (x, y) \in \mathbb{H} \times \mathbb{R}. \tag{3.1}$$

To estimate it, we have access to a data sample $\{(X_i, Y_i), i = 1, ..., n\}$ distributed like the couple (X, Y).

Remark 3: We give here some precision about existence and uniqueness of the function $F^x(y)$. For all $y \in \mathbb{R}$, the random variable $\mathbb{P}(Y \leq y|X)$ is $\sigma(X)$-measurable (where $\sigma(X)$ is the σ-field generated by X). Hence, there exists a function $\Phi : \mathbb{H} \times \mathbb{R} \to [0, 1]$ such that:

– $x \mapsto \Phi(x, y)$ is measurable;

– $\Phi(X, y) = F^X(y)$ p.s.

The quantity $\Phi(x, y)$ is uniquely defined as soon as x is in the support of X.

It can be shown that the Dynkin's $\pi - \lambda$ theorem implies that the map $y \mapsto \Phi(x, y)$ is measurable for all x in the support of the random variable X, the proof is very similar to the proof of Fubini's theorem.

When x is not in the support of X we can set the values of $\Phi(x, y)$ in the most suitable way. For instance, if, Φ is continuous on $\text{int}(\text{Supp}(X)) \times \mathbb{R}$ [1] we can extend it by continuity on $\overline{\text{Supp}}(X) \times \mathbb{R}$ [2]. And if the support of X is convex (which is the case for instance when X is a Gaussian process since the support of X is the closure of a vector space), we can define, for all $(x, y) \in \mathbb{H} \times \mathbb{R}$, $\Phi(x, y) = \Phi(\pi x, y)$ where π is the projection on the support of X.

We denote $F^x(y) = \Phi(x, y)$.

3.1.1 Motivation

The pioneering works on conditional distribution when the covariate is functional are the one of Ferraty and Vieu (2002) and Ferraty, Laksaci, and Vieu (2006), completed by Ferraty, Laksaci, Tadj, et al. (2010). Kernel estimators, which depend on a smoothness parameter, the so-called bandwidth, are built to address several estimation problems: regression function, conditional c.d.f., conditional density and its derivatives, conditional hazard rate, conditional mode and quantiles. A lot of research has then been carried out to extend or adapt the previous procedures to various statistical models. For instance, the estimation of the regression function is studied by Rachdi and Vieu (2007), Ferraty, Mas, and Vieu (2007), and Dabo-Niang and Rhomari (2009). The case of dependent data is the subject of the works of Masry (2005), Aspirot, Bertin, and Perera (2009), Laib and Louani (2010), and Dabo-Niang, Kaid, and Laksaci (2012) under several assumptions (α-mixing, ergodic or non-stationary processes). Robust versions of the previous strategies are proposed by Crambes, Delsol, and Laksaci, 2008; Azzedine, Laksaci, and Ould-Saïd, 2008; Gheriballah, Laksaci, and Sekkal, 2013. Most of this literature focuses on asymptotic results (almost-complete convergence, asymptotic normality,...). Bias-variance decompositions are provided. However, few papers tackle the problem of bandwidth selection: Rachdi and Vieu, 2007 and Benhenni, Ferraty, et al., 2007 have studied global or local cross-validation procedures which are shown to be asymptotically optimal in regression contexts. Recently, a Bayesian criterion has been investigated from a numerical point of view by Shang, 2013.

The main goal is to define a fully data-driven selection rule for the bandwidth h, which satisfies nonasymptotic adaptive results. The criterion we propose draws inspiration from both the so-called Lepski method (see the recent paper of Goldenshluger and Lepski 2011) and model selection tools. We also study minimax rates under several assumptions on the decreasing rate of the small ball probability $\varphi(h) = \mathbb{P}(d(X, 0) \leq h)$ to 0. The case of projection semi-norms, which have been widely studied is also considered.

1. $\text{int}(\text{Supp}(X))$: interior of the support of X
2. $\overline{\text{Supp}}(X)$: closure of the support of X

3.1.2 Definition of the estimator with a fixed bandwidth

To estimate the c.d.f. defined by (3.1), we consider

$$\widehat{F}_h^x(y) := \sum_{i=1}^n W_h^{(i)}(x)\mathbf{1}_{\{Y_i \le y\}} \text{ where } W_h^{(i)}(x) := \frac{K_h(d(X_i,x))}{\sum_{j=1}^n K_h(d(X_j,x))}, \tag{3.2}$$

for any $(x,y) \in \mathbb{H} \times \mathbb{R}$, with d a general semi-metric on the Hilbert space $\mathbb{H}$, $K_h : t \mapsto K(t/h)/h$, for K a kernel function (that is $\int_{\mathbb{R}} K(t)dt = 1$) and h a parameter to be chosen, the so-called bandwidth. We focus on the metric associated to the norm of the Hilbert space

$$d(x,x') := \|x - x'\|, \quad x, x' \in \mathbb{H}. \tag{3.3}$$

It has been pointed out (see e.g. Ferraty, Laksaci, and Vieu 2006) that, when the data is high or infinite-dimensional, the convergence rate is quite slow with this choice of semi-norm. To bypass this *curse of dimensionality* problem, some researchers (see e.g. Masry 2005; Ferraty, Laksaci, and Vieu 2006; Geenens 2011) have suggested replacing the norm $\|\cdot\|$ in the definition of the estimator (3.2) by a semi-norm. The case of projection semi-norms has received particular attention. In that case the estimator can be redefined this way

$$\widehat{F}_{h,p}^x(y) := \sum_{i=1}^n W_{h,p}^{(i)}(x)\mathbf{1}_{\{Y_i \le y\}} \text{ with } W_{h,p}^{(i)}(x) := \frac{K_h(d_p(X_i,x))}{\sum_{j=1}^n K_h(d_p(X_j,x))}, \tag{3.4}$$

where $d_p^2(x,x') := \sum_{j=1}^p \langle x - x', e_j\rangle^2$ and $(e_j)_{j\ge1}$ is a basis of $\mathbb{H}$. Defining this estimator amounts to project the data into a p-dimensional space.

3.1.3 Considered risks

We consider two types of risks for the estimation of $(x,y) \mapsto F^x(y)$. Both are mean integrated squared error with respect to the response variable y.

The first criterion is a *pointwise risk* in x, integrated in y:

$$\mathbb{E}\left[\|\widehat{F}_h^{x_0} - F^{x_0}\|_D^2\right],$$

for a fixed $x_0 \in \mathbb{H}$, D a compact subset of $\mathbb{R}$ and

$$\|f\|_D^2 := \int_D f(t)^2 dt,$$

keeping in mind that the Hilbert norm of $\mathbb{H}$ is $\|.\|$. We also denote by $|D| := \int_D dt$ the Lebesgue measure of the set D.

Next, we introduce a second criterion, which is an *integrated risk* with respect to the product of the Lebesgue measure on $\mathbb{R}$ and the probability measure $\mathbb{P}_X$ of X, defined by

$$\mathbb{E}\left[\|\widehat{F}_h^{X'} - F^{X'}\|_D^2 \mathbf{1}_B(X')\right] = \mathbb{E}\left[\int_D \int_B \left(\widehat{F}_h^x(y) - F^x(y)\right)^2 dy d\mathbb{P}_X(x)\right], \tag{3.5}$$

where X' is a copy of X independent of the data sample and B is a subset of $\mathbb{H}$.

The motivation for studying the two risks is twofold. First, in practice, we can either be interested in the estimation of $F^{X_{n+1}}$ where X_{n+1} is a copy of X independent of the sample or we can be interested in estimating the c.d.f conditionally to $X = x_0$ where x_0 is a point chosen in advance. For instance, for the spectrometric dataset, presented in Section 3.6.4, the aim is to understand the repartition of the fat content Y given a new spectrometric curve X' which, as we can assume logically, follows the same distribution as X. Hence, in that context, studying the distribution of Y given $X = x_0$, where x_0 is a deterministic curve of $\mathbb{H}$ has a limited practical interest. Such an approach is rather classical in functional linear regression (Ramsay and Silverman, 2005; Cardot, Ferraty, and Sarda, 1999) where either prediction error on random curves (Crambes, Kneip, and Sarda, 2009) or prediction error over a fixed curve (Cai and Hall, 2006) are considered. Second, integrated risks have been relatively unexplored in non-parametric functional data analysis. Indeed, there is no measure universally accepted as the Lebesgue measure in finite-dimensional setting (see e.g. Delaigle and Hall 2010; Dabo-Niang and Yao 2013). The only measure at hand is the probability measure of X.

3.1.4 Organisation of the chapter

In Section 3.2, we provide a bias-variance decomposition of the estimator (3.2) in terms of two criteria, a pointwise and an integrated risk. The bandwidth h is shown to influence significantly the quality of estimation. In Section 3.3, we define a bandwidth selection criterion achieving the best bias-variance trade-off. Rates of convergence of the resulting estimator are computed in Section 3.4. Consistently with the previous works, the rates we obtain are quite slow, but we also prove lower bounds showing that these rates are optimal. The results are also shown to be coherent with lower bounds computed by Mas (2012) for the estimation of the regression function. Properties of the estimator defined with a projection semi-metric are investigated in Section 3.5. We show that this method does not improve the convergence rates of the Nadaraya-Watson estimator since the lower bounds are still valid. In order to understand what is going on, we briefly study a bias-variance decomposition of the risk of this estimator. Finally, the proofs are gathered in Section 3.7.

3.2 Integrated and pointwise risk of an estimator with fixed bandwidth

3.2.1 Assumptions

Hereafter, we denote by φ^x the shifted small ball probability:

$$\varphi^x(h) = \mathbb{P}(\|X - x\| \leq h), \quad h > 0, \quad x \in \mathbb{H}.$$

We write $\varphi(h)$ instead of $\varphi^0(h)$. If X' is a random variable, $\varphi^{X'}$ is the conditional small ball probability: $\varphi^{X'}(h) = \mathbb{P}_{X'}(\|X - X'\| \leq h)$, where hereafter the notation $\mathbb{P}_{X'}$ (resp. $\mathbb{E}_{X'}$, $\mathrm{Var}_{X'}$) stands for the conditional probability (resp. expectation, variance) given X'. For simplicity, we assume that the curve X is centred. We also consider the following assumptions. The first one is related to the choice of the kernel, the two following are

regularity assumptions for the function to estimate and the process X.

H_K The kernel K is of type I (Ferraty and Vieu, 2006) i.e. its support is in $[0,1]$ and there exist two constants $c_K, C_K > 0$ such that

$$c_K \mathbf{1}_{[0,1]} \leq K \leq C_K \mathbf{1}_{[0,1]}.$$

H_F There exists $\beta \in]0,1[$ such that F^x belongs to the functional space $\mathcal{F}_\beta$, the class of the maps $(x,y) \in \mathbb{H} \times \mathbb{R} \mapsto F^x(y)$ such that:
- for all $x \in \mathbb{H}$, F^x is a c.d.f;
- there exists a constant $C_D > 0$ such that, for all $x, x' \in \mathbb{H}$

$$\|F^x - F^{x'}\|_D \leq C_D \|x - x'\|^\beta.$$

H_φ There exist two constants $c_\varphi, C_\varphi > 0$ such that for all $h \in \mathbb{R}$, for all $x_0 \in B$,

$$c_\varphi \varphi(h) \leq \varphi^{x_0}(h) \leq C_\varphi \varphi(h).$$

Assumption H_K is quite classical in kernel methods for functional data (see Ferraty, Laksaci, and Vieu 2006; Burba, Ferraty, and Vieu 2009; Ferraty, Laksaci, Tadj, et al. 2010). We are aware that this is a strong assumption but alleviate it in a functional data context requires a lot of technical difficulties and it is still, to our knowledge, an open problem.

Assumption H_F is an Hölder-type regularity condition on the map $x \mapsto F^x$. This type of condition is natural in kernel estimation. It is very similar to Assumption (H2) of Ferraty, Laksaci, and Vieu (2006) or Assumption (H2') of Ferraty, Laksaci, Tadj, et al. (2010). Note, however, that, since both considered risks are integrated with respect to y, no regularity condition on the map $y \mapsto F^x(y)$ is required here. A similar phenomenon appears for the estimation of the c.d.f when the covariate is real: for instance, the convergence rate given by Brunel, Comte, and Lacour (2010, Corollary 1) only depends on the regularity of F with respect to x. For instance if $Y = m(X) + \varepsilon$, where m is a function and ε a real random variable independent of X, we have $F^x(y) = \mathbb{P}(m(x) + \varepsilon \leq y) = F_\varepsilon(y - m(x))$ where F_ε is the cumulative distribution function of ε. Hence the regularity properties of F^x with respect to x depend on the regularity of both m and F_ε: if F_ε is β_1-hölderian and m is β_2-hölderian with $\beta = \beta_1\beta_2$, we can verify easily that H_F is verified.

Assumption H_φ is very similar to assumptions made by Ferraty, Laksaci, and Vieu 2006; Burba, Ferraty, and Vieu 2009; Ferraty, Laksaci, Tadj, et al. 2010. This condition H_φ is reasonable, since the class of Gaussian processes fulfill it provided that B is a bounded subset of $\mathbb{H}$. Indeed the upper bound is verified with $C_\varphi = 1$ thanks to Anderson's Inequality (Anderson, 1955) (see also Li and Shao 2001, Theorem 2.13 or Hoffmann-Jørgensen, Shepp, and Dudley 1979, Theorem 2.1, p.322) and from Hoffmann-Jørgensen, Shepp, and Dudley (1979, Theorem 2.1, p.322) we know that the lower bound is verified with $c_\varphi := e^{-R^2/2}$ where $R := \max\{\|x\|,\ x \in B\}$.

3.2.2 Upper bound

Under the assumptions above we are able to obtain a non-asymptotic upper bound for the risk:

Theorem 3. *Suppose assumptions H_K and H_F are fulfilled. Let $h > 0$ be fixed.*

(i) For all $x_0 \in \mathbb{H}$ we have

$$\mathbb{E}\left[\left\|\widehat{F}_h^{x_0} - F^{x_0}\right\|_D^2\right] \leq C\left(h^{2\beta} + \frac{1}{n\varphi^{x_0}(h)}\right), \tag{3.6}$$

where $C > 0$ only depends on c_K, C_K, $|D|$ and C_D.

(ii) If, in addition, Assumption H_φ is fulfilled,

$$\mathbb{E}\left[\|\widehat{F}_h^{X'} - F^{X'}\|_D^2 \mathbf{1}_B(X')\right] \leq C\left(h^{2\beta} + \frac{1}{n\varphi(h)}\right), \tag{3.7}$$

where $C > 0$ only depends on c_K, C_K, c_φ, C_φ, $|D|$ and C_D.

The first term of the right-hand-side of inequalities (3.6) and (3.7) corresponds to a bias term, and the second is a variance term, which increases when h goes to 0 (since $\varphi^{x_0}(h)$ and $\varphi(h)$ decrease to 0 when $h \to 0$). Note that the upper bounds are very similar to the results of Ferraty, Laksaci, and Vieu (2006, Theorem 3.1) and Ferraty, Laksaci, Tadj, et al. (2010, Corollary 3). However, we do not have an extra-$\ln n$ factor in the variance term.

We deduce from Theorem 3 that the usual bias-variance trade-off must be done if one wants to choose h in a family of possible bandwidths. The ideal compromise h^* is called the oracle, and is defined by

$$h^* = \arg\min_h \mathbb{E}\left[\left\|\widehat{F}_h^{X'} - F^{X'}\right\|_D^2 \mathbf{1}_B(X')\right]. \tag{3.8}$$

It cannot be used as an estimator since it both depends on the unknown regularity index β of F and on the rate of decrease of the small ball probability $\varphi(h)$ of X to 0. The challenge is to propose a fully data-driven method to perform the trade-off.

3.3 Adaptive estimation

In this section, we focus on the integrated risk. We refer to Remark 4 below for the extension of the results for the pointwise criterion.

3.3.1 Bandwidth selection

We have at our disposal the estimators $\widehat{F}_h$ defined by (3.2) for any $h > 0$. Let $\mathcal{H}_n$ be a finite collection of bandwidths, with cardinality depending on n and properties precised below. For any $h \in \mathcal{H}_n$, an empirical version for the small ball probability $\varphi(h) = \mathbb{P}(\|X\| \leq h)$ is

$$\widehat{\varphi}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\|X_i\| \leq h\}}. \tag{3.9}$$

For any $h \in \mathcal{H}_n$, we define

$$\widehat{A}(h) = \max_{h' \in \mathcal{H}_n} \left(\left\| \widehat{F}^{X'}_{h'} - \widehat{F}^{X'}_{h \vee h'} \right\|_D^2 - \widehat{V}(h') \right)_+, \quad \widehat{V}(h) = \begin{cases} \kappa \dfrac{3\ln(n)}{2n\widehat{\varphi}(h)} \text{ if } \widehat{\varphi}(h) \neq 0 \\ +\infty \text{ otherwise,} \end{cases} \tag{3.10}$$

where κ is a constant specified in the proofs which depends neither on h, nor on n, nor on $F^{X'}$. The quantity $\widehat{V}(h)$ is an estimate of the upper bound for the variance term (see (3.7)) and $\widehat{A}(h)$ is proved to be an approximation of the bias term (see Lemma 12). This motivates the following choice of the bandwidth:

$$\widehat{h} = \text{argmin}_{h \in \mathcal{H}_n} \left\{ \widehat{A}(h) + \widehat{V}(h) \right\}.$$

The selected estimator is $\widehat{F}_{\widehat{h}}$.

This selection rule is inspired both by the recent version of the so-called Lepski method (see Goldenshluger and Lepski 2011) and by model selection tools. The main idea is to estimate the bias term by looking at several estimators. Goldenshluger and Lepski (2011) propose to first define "intermediate" estimates $\widehat{F}^{X'}_{h,h'}$ $(h, h' \in \mathcal{H}_n)$, based on a convolution product of the kernel with the estimators with fixed bandwidths. However, this can only be done when the bias of the estimator is written as the convolution product of the kernel with the target function. Since it is not the case in our problem, we perform the bandwidth selection with $\widehat{F}^{X'}_{h,h'} = \widehat{F}^{X'}_{h \vee h'}$ in (3.10). This is analogous to the procedure proposed by Chagny (2013b) or Comte and Johannes (2012) for model selection purpose. Thus, $V(h)$ can also be seen as a penalty term. We also refer to the phD of Chagny (2013a, p.170) for technical details leading to this choice. Finally, let us notice that a criterion based on the maximum $h \vee h'$ also appears in Kerkyacharian, Lepski, and Picard, 2001, and more recently, similar ideas are used in Goldenshluger and Lepski, 2013b.

3.3.2 Theoretical results

To prove our main results, we consider the following hypothesis, in addition to the assumptions defined in Section 3.2.1.

H_b The collection $\mathcal{H}_n$ of bandwidths is such that:

H_{b1} its cardinality is bounded by n,

H_{b2} for any $h \in \mathcal{H}_n$, $\varphi(h) \geq C_0 \ln(n)/n$, where $C_0 > 0$ is a purely numerical constant (specified in the proofs).

Assumption H_{b1} fixes the size of the bandwidth collection: compared to the assumptions of Goldenshluger and Lepski (2011), we consider a discrete set and not an interval, which permits to use the classical tools of model selection theory in the proofs. We now state the following result.

In practice, it is impossible to verify Assumption H_{b2} since the function φ and the constant C_0 are unknown. However, this difficulty can be circumvented by introducing a random collection of bandwidths $\widehat{\mathcal{H}}_n$ verifying, for all $h \in \widehat{\mathcal{H}}_n$, $\widehat{\varphi}(h) \geq 2\widehat{C}_0 \ln(n)/n$ where $\widehat{\varphi}$ and $\widehat{C}_0$ are some estimators of φ (see Equation (3.9)) and C_0. However, since it would complicate the understandability of proofs, we choose to keep Assumption H_{b_2}.

Theorem 4. *Assume H_K, H_φ, H_F, H_b and that $n \geq 3$. There exist two constants $c, C > 0$ depending on c_K, C_K, c_φ, C_φ, $|D|$, C_D such that*

$$\sup_{F \in \mathcal{F}_\beta} \mathbb{E}\left[\left\|\widehat{F}^{X'}_{\widehat{h}} - F^{X'}\right\|_D^2 \mathbf{1}_B(X')\right] \leq c \min_{h \in \mathcal{H}_n}\left\{h^{2\beta} + \frac{\ln(n)}{n\varphi(h)}\right\} + \frac{C}{n}. \tag{3.11}$$

The optimal bias-variance compromise is reached by the estimator, which is thus adaptive with respect to the unknown smoothness of the target function F. Even if Equation (3.11) is not strictly speaking an oracle-inequality, it shows that the selected bandwidth $\widehat{h}$ is performing as well as the unknown oracle h^* defined in (3.8), up to the multiplicative constant c, up to a remainding term of order $1/n$ which is negligible, and up to the $\ln(n)$ factor. This extra-quantity also appears in the term $V(h)$. Usually, when integrated risks are considered, there is no logarithmic loss due to adaptation (Brunel, Comte, and Lacour, 2010; Goldenshluger and Lepski, 2011), then further work is necessary to see if it is possible to improve Inequality (3.11). Nevertheless, we prove in Section 3.4 that it does not affect significantly the convergence rates of the estimator which is optimal in the minimax sense in most of the cases.

The proof of Theorem 4 is mainly based on model selection tools, specifically concentration inequalities. A specific difficulty comes from the fact that the variance term in (3.7) depends on the unknown distribution of X, through its small ball probability. Thus, the penalty term $V(h) = \kappa \ln(n)/(n\varphi(h))$, which may have been classically defined cannot be used in practice. This explains why we plug (3.9), an estimate for $\varphi(h)$ in $\widehat{V}(h)$. However, for the sake of clarity, we begin the proof by establishing the result with $\widehat{V}(h)$ replaced by its theoretical counterpart $V(h) = \kappa \ln(n)/(n\varphi(h))$.

Remark 4: We could build an adaptive estimator for the pointwise risk. To do so, replace $\widehat{\varphi}(h)$ in (3.10) by $\widehat{\varphi}^{x_0}(h) = \sum_{i=1}^n \mathbf{1}_{\{\|X_i - x_0\|) \leq h\}}/n$ and X' by x_0 in the definition of $\widehat{A}(h)$.

3.4 Minimax rates

In this section, we compute the convergence rate of the oracle $\widehat{F}_{h^*}$ with h^* defined by (3.8), the rate of the selected estimator $\widehat{F}_{\widehat{h}}$, and prove lower bounds for the conditional c.d.f. estimation problem under various assumptions on the rate of decrease of the small ball probability of the covariate X.

3.4.1 Small ball probabilities

The computation of the oracle h^*, as well as the computation of the minimum in the right-hand-side of (3.11) require to fix conditions on the rate of decrease of the small ball probability $\varphi(h)$. The choice of the assumptions is crucial and determines the rates of convergence to zero of our estimators. We consider in the sequel one of the three following hypothesis which allow to understand how the small ball probability decay influences the rates (see Section 3.4.2) and which are frequently used in the literature. We describe below large class of processes for which they are fulfilled.

H_X There exist some constants $c_1, C_1 > 0$ such that $\varphi^{x_0}(h)$ satisfies one of the following three assumptions, for any $h > 0$.

$H_{X,L}$ There exist some constants $\gamma_1, \gamma_2 \in \mathbb{R}$, and $\alpha > 0$ such that $c_1 h^{\gamma_1} \exp(-c_2 h^{-\alpha}) \leq \varphi^{x_0}(h) \leq C_1 h^{\gamma_2} \exp(-c_2 h^{-\alpha})$;

$H_{X,M}$ There exist some constants $\gamma_1, \gamma_2 \in \mathbb{R}$, and $\alpha > 1$, such that $c_1 h^{\gamma_1} \exp(-c_2 \ln^{\alpha}(1/h)) \leq \varphi^{x_0}(h) \leq C_1 h^{\gamma_2} \exp(-c_2 \ln^{\alpha}(1/h))$;

$H_{X,F}$ There exists a constant $\gamma > 0$, such that $c_1 h^{\gamma} \leq \varphi^{x_0}(h) \leq C_1 h^{\gamma}$,

where we set $x_0 = 0$ if we consider the integrated risk.

Such inequalities are heavily connected with the rate of decrease of the eigenvalues of the covariance operator $\Gamma : f \in \mathbb{H} \mapsto \Gamma f \in \mathbb{H}$ with $\Gamma f(s) = \langle f, \mathrm{Cov}(X_\cdot, X_s) \rangle$. Recall the Karhunen-Loève decomposition of the process X, which can be written

$$X = \sum_{j \geq 1} \sqrt{\lambda_j} \eta_j \psi_j, \tag{3.12}$$

where $(\eta_j)_{j \geq 1}$ are uncorrelated real-random variables, $(\lambda_j)_{j \geq 1}$ is a non-increasing sequence of positive numbers (the eigenvalues of Γ) and $(\psi_j)_{j \geq 1}$ an orthonormal basis of $\mathbb{H}$. When X really lies in an infinite dimensional space, the set $\{j \geq 1,\ \lambda_j > 0\}$ is infinite, and under mild assumptions on the distribution of X, it is known that $\varphi(h)$ decreases faster than any polynomial of h (see e.g. Mas 2012, Corollary 1, p.10). This is the case in Assumptions $H_{X,L}$ and $H_{X,M}$. Moreover, the faster the decay of the eigenvalues is, the more the data are concentrated close to a finite dimensional space, and the slower $\varphi(h)$ decreases.

For example, when X is a Gaussian process with eigenvalues $(\lambda_j)_j$ such that $cj^{-2a} \leq \lambda_j \leq Cj^{-2a}$, $a \geq 1/2$ $(c, C > 0)$, Assumption $H_{X,L}$ is satisfied with $\gamma_1 = \gamma_2 = (3 - a)/(2a - 1)$, $c_2 = a(2a/(2a-1))^{1/(2a-1)}$ and $\alpha = 1/(a - 1/2)$ (Hoffmann-Jørgensen, Shepp, and Dudley 1979, Theorem 4.4 and example 4.5, p.333-334). This classical situation of such polynomial decay covers the example of the Brownian motion, with $a = 1$ (see Ash and Gardner 1975 or Section 1.1.2, p.6). More generally, if X is defined by a random series $X = \sum_{j \geq 1} j^{-2a} Z_j$, for variables Z_i with a c.d.f. regularly varying at 0 with positive index, one can also define γ_1, γ_2, and α such that $H_{X,L}$ is fulfilled (see Dunker, Lifshits, and Linde 1998, Proposition 4.1 p.11 and also Mas 2012, (19) p.9). The second case $H_{X,M}$ typically happens when the eigenvalues of the covariance operator exponentially decrease (see Dunker, Lifshits, and Linde 1998, Proposition 4.3 p.12). In the case of a Gaussian process with $c \exp(-2j)/j \leq \lambda_j \leq C \exp(-2j)/j$, we have $c_2 = 1/2$ and $\alpha = 2$ in $H_{X,M}$ (Hoffmann-Jørgensen, Shepp, and Dudley, 1979, Theorem 4.4 and example 4.7, pp. 333 and 336).

Finally, it also results of the above considerations that $H_{X,F}$ only covers the case of finite dimensional processes (the set $\{j,\ \lambda_j > 0\}$ is finite, that is the operator Γ has a finite rank). This is the extreme case of $H_{X,M}$ (with $\alpha = 1$, $\gamma_1 = \gamma_2 = 0$). Nevertheless, even if our main purpose is to study functional data, the motivation to keep this case is twofold. First, we show below that our estimation method allows to recover the classical rates (upper and lower bounds) obtained for c.d.f. estimation with multivariate covariates. Then, processes which fulfill $H_{X,F}$ can still be considered as functional data since the finite space to whom X belongs is unknown for the statistician.

		$H_{X,L}$ (lower rate)	$H_{X,M}$ (medium rate)	$H_{X,F}$ (fast rate)
(a)	Rates for $\widehat{F}_{h^*}$ (upper bounds)	$(\ln(n))^{-2\beta/\alpha}$	$\exp\left(-\frac{2\beta}{c_2^{1/\alpha}}\ln^{1/\alpha}(n)\right)$	$n^{-\frac{2\beta}{2\beta+\gamma}}$
(b)	Rates for $\widehat{F}_{\widehat{h}}$ (upper bounds)	$(\ln(n))^{-2\beta/\alpha}$	$\exp\left(-\frac{2\beta}{c_2^{1/\alpha}}\ln^{1/\alpha}(n)\right)$	$\left(\frac{n}{\ln(n)}\right)^{-\frac{2\beta}{2\beta+\gamma}}$
(c)	Minimax risk (lower bounds)	$(\ln(n))^{-2\beta/\alpha}$	$\exp\left(-\frac{2\beta}{c_2^{1/\alpha}}\ln^{1/\alpha}(n)\right)$	$n^{-\frac{2\beta}{2\beta+\gamma}}$

Table 3.1: Rates of convergence of the oracle estimator (line (a)) and the adaptive estimator (line (b)). Minimax lower bounds (line (c)).

3.4.2 Convergence rates of kernel estimators

We now compute the upper bounds for the pointwise and integrated risks of the estimators, under the previous regularity assumptions.

Theorem 5. *(a) Under the assumptions of Theorem 3, the convergence rates of the pointwise risk* $\mathbb{E}[\|\widehat{F}_{h^*}^{x_0} - F^{x_0}\|^2]$, *and the integrated risk* $\mathbb{E}[\|\widehat{F}_{h^*}^{X'} - F^{X'}\|_D^2]$ *of the oracle* $\widehat{F}_{h^*}$ *are given in Table 3.1, line (a).*

(b) Under the assumptions of Theorem 4, the convergence rates of the integrated risk $\mathbb{E}[\|\widehat{F}_{\widehat{h}}^{X'} - F^{X'}\|_D^2 \mathbf{1}_B(X')]$ *of the estimator* $\widehat{F}_{\widehat{h}}$ *are given in Table 3.1, line (b).*

For both cases, the upper bounds are given up to a multiplicative constant, and for the different cases $H_{X,L}$, $H_{X,M}$, *and* $H_{X,F}$.

Let us comment the results. The faster the small ball probability decreases (that is the less concentrated the measure of X is), the slower the rate of convergence of the estimator is. In the generic case of a process X which satisfies $H_{X,L}$, the rates are logarithmic, which is not surprising. It reflects the "curse of dimensionality" which affects the functional data. Similar rates are obtained by Ferraty, Laksaci, and Vieu (2006) (section 5.3) in the same framework, and by Mas (2012) for regression estimation (section 2.3.1). However, we show that the results can be improved when the process X is more regular, although still infinite dimensional. Under Assumption $H_{X,M}$, the rates we compute have the property to decrease faster than any logarithmic function. Assumption $H_{X,F}$ is the only one which yields to the faster rate, that is the polynomial one.

Remark 5: We thus obtain various rates, depending on the regularity assumptions on X.

This phenomenom also occurs in a deconvolution model: the rates for the kernel estimators are logarithmic if the noise is "supersmooth" and the signal to recover "ordinarysmooth", but can be improved by considering the case of a "supersmooth" signal, to recover at least rates which are intermediate between logarithmic and polynomial (see e.g. Lacour 2006; Comte and Lacour 2010).

We have already noticed that our adaptive procedure leads to the loss of a logarithm factor (see the comments following Theorem 4). Nevertheless, by comparing line (a) to line (b) in Table 3.1, we obtain that the adaptive estimator still achieves the oracle rate if $H_{X,L}$ or $H_{X,M}$ are fulfilled. The loss is actually negligible with respect to the rates.

3.4.3 Lower bounds

We now establish lower bounds for the risks under mild additional assumptions, showing that the estimators suggested above attain the optimal rates of convergence in a minimax sense over the class of conditional c.d.f. $\mathcal{F}_\beta$ (defined in Section 3.2.1). The results for the integrated risk are obtained through non-straightforward extensions of the pointwise case.

Theorem 6. *Suppose that H_X is fulfilled, and that $n \geq 3$.*

(i) The minimax risk $\inf_{\widehat{F}} \sup_{F \in \mathcal{F}_\beta} \mathbb{E}_F[\|\widehat{F}^{x_0} - F^{x_0}\|^2]$ is lower bounded by a quantity proportional to the ones in line (c) in Table 3.1.

(ii) Assume moreover that B contains the ball $\{x \in \mathbb{H},\ \|x\| \leq \rho\}$ where $\rho > 0$ is a constant to be specified in the proof, and that there exist two constants $c_2, C_2 > 0$ such that, for all $h > 0$, for all $x \in B$,

$$\varphi(h) > 0 \text{ and } c_2\varphi(h) \leq \varphi^x(h) \leq C_2\varphi(h). \tag{3.13}$$

Then the minimax risk $\inf_{\widehat{F}} \sup_{F \in \mathcal{F}_\beta} \mathbb{E}_F[\|\widehat{F}^{X'} - F^{X'}\|_D^2 \mathbf{1}_B(X')]$ is also lower bounded by a quantity proportional to the ones in line (c) in Table 3.1.

For both cases, the infimum is taken over all possible estimators obtained with the data-sample $(X_i, Y_i)_{i=1,\dots,n}$. In (i)*, $\mathbb{E}_F$ is the expectation with respect to the law of $\{(X_i, Y_i), i = 1, \dots, n\}$ and in* (ii)*, $\mathbb{E}_F$ is the expectation with respect to the law of $\{\{(X_i, Y_i), i = 1, \dots, n\}, X'\}$ when, for all $i = 1, \dots, n$, for all $x \in \mathbb{H}$, the conditional c.d.f. of Y_i given $X_i = x$ is F^x.*

Theorem 6 proves that the upper bounds of Theorem 5 cannot be improved, not only among kernel estimators but also among all estimators, under assumptions $H_{X,L}$ and $H_{X,M}$. The estimator $\widehat{F}_{\widehat{h}}$ is thus both adaptive optimal in the oracle and in the minimax senses.

The computations are new for conditional c.d.f. estimation with a functional covariate. Under $H_{X,F}$, with $\gamma = 1$, the lower bounds we obtain are consistent with Theorem 2 of Brunel, Comte, and Lacour (2010) or Proposition 4.1 of Plancade (2013) for c.d.f. estimation with a one-dimensional covariate, over Besov balls. In the functional framework, the results can only be brought close to those of Mas (2012) (Theorem 3) for regression estimation.

3.5 Impact of the projection of the data on finite-dimensional spaces

We have seen in Section 3.4.2 that, when X lies in an infinite dimensional space (assumptions $H_{X,M}$ and $H_{X,L}$), the rates of convergence are slow. This "curse of dimensionality" phenomenon is well known in kernel estimation for high or infinite dimensional datasets. The introduction of the projection semi-metrics d_p, leading to the estimators (3.4), has thus been proposed in order to circumvent this problem. Defining such estimators amounts to project the data into a p-dimensional space. Indeed, this permits to address the problem of variance reduction since $\varphi_p(h) := \mathbb{P}(d_p(x,0) \leq h) \sim_{h\to 0} C(p)h^p$ and then the variance is of order $1/(nh^p)$. Notice that, even if the variance order of magnitude are the same, the situation here is different from Assumption $H_{X,F}$ with $\gamma = p$: $H_{X,F}$ amounts to suppose that the curve X lies a.s. in an unknown finite-dimensional space (see Section 3.4.1) whereas, here, the data are projected into a finite-dimensional space but may lie in an infinite-dimensional space.

A first thing we can say is that, under our regularity assumption H_F, Theorem 6 remains true and the convergence rate of the risk of $\widehat{F}_{h,p}$ cannot be better than the lower bounds given in Table 3.1, line (c). This implies that, in our setting, the estimator $\widehat{F}_{h,p}$ cannot converge at significantly better rates than our adaptive estimator $\widehat{F}_{\widehat{h}}$ even if the couple of parameters (p,h) is well chosen. Precisely, as shown in Proposition 2 below, project data also adds an additional bias term which compensates for the decrease of the variance.

3.5.1 Assumptions

In order to state the result, we need the following assumptions.

H'_φ There exist two constants $c_\varphi, C_\varphi > 0$ such that for all $h \in \mathbb{R}$, for all $p \in \mathbb{N}^*$,

$$c_\varphi \varphi_p(h)\mathbf{1}_B(X') \leq \varphi_p^{X'}(h)\mathbf{1}_B(X') \leq C_\varphi \varphi_p(h)\mathbf{1}_B(X') \text{ a.s.},$$

where X' is an independent copy of X and $\varphi_p^{X'}(h) := \mathbb{P}_{X'}(d_p(X,X') \leq h)$.

H_ξ Let $\xi_j := \langle X, e_j\rangle/\sigma_j$ where $\sigma_j := \mathrm{Var}(\langle X, e_j\rangle)$. One of the two following assumptions is verified:

H_ξ^{ind} the sequence of random variables $(\xi_j)_{j\geq 1}$ is independent and there exists a constant C_ξ such that, for all $j \geq 1$

$$\mathbb{E}\left[\xi_j^\beta\right] \leq C_\xi;$$

H_ξ^b there exists a constant C_ξ such that, for all $j \geq 1$,

$$|\xi_j| \leq C_\xi \text{ a.s.}$$

Remark that Assumption H'_φ is the equivalent of Assumption H_φ replacing d by d_p. If X is a Gaussian process, the vector $(\langle X, e_1\rangle, \ldots, \langle X, e_p\rangle)$ is a Gaussian vector and Assumption H_φ is also verified provided that B is bounded. Assumption H_ξ^{ind} is true if X is a Gaussian process and $(e_j)_{j\geq 1}$ is the Karhunen-Loève basis of X (see (3.12) above, and also Ash and

Gardner 1975) and Assumption H_ξ^b is equivalent to suppose that X is bounded a.s. We are aware that both assumptions H_ξ^{ind} and H_ξ^b are strong since in most cases the Karhunen-Loève basis is unknown. We give here Proposition 2 below in the only aim of better understanding the bias-variance decomposition of the risk when the data are projected. A further study would be needed to obtain weaker assumptions but this is beyond the scope of this work.

3.5.2 Upper bound

Proposition 2. *Suppose assumptions H_K, H_F and H_ξ are fulfilled. Let $h > 0$ and $p \in \mathbb{N}^*$ be fixed.*

(i) For all $x_0 \in \mathbb{H}$ we have

$$\mathbb{E}\left[\left\|\widehat{F}_{h,p}^{x_0} - F^{x_0}\right\|_D^2\right] \leq C\left(h^{2\beta} + \left(\sum_{j>p}\sigma_j^2\right)^\beta + \left(\sum_{j>p}\langle x_0, e_j\rangle^2\right)^\beta + \frac{1}{n\varphi_p^{x_0}(h)}\right), \tag{3.14}$$

where $C > 0$ only depends on C_ξ, β, c_K, C_K, $|D|$ and C_D.

(ii) If, in addition, Assumption H_φ is fulfilled,

$$\mathbb{E}\left[\|\widehat{F}_{h,p}^{X'} - F^{X'}\|_D^2 \mathbf{1}_B(X')\right] \leq C\left(h^{2\beta} + \left(\sum_{j>p}\sigma_j^2\right)^\beta + \frac{1}{n\varphi_p(h)}\right), \tag{3.15}$$

where $C > 0$ only depends on C_ξ, β, c_K, C_K, c_φ, C_φ, $|D|$ and C_D.

We have additional bias terms compared to Ferraty, Laksaci, and Vieu (2006) and Ferraty, Laksaci, Tadj, et al. (2010). This is due to the fact that our regularity assumption H_F (see Section 3.2.1) is here different from Assumption (H2) of Ferraty, Laksaci, and Vieu (2006) or Assumption (H2') of Ferraty, Laksaci, Tadj, et al. (2010). Our assumption is expressed with the norm of $\mathbb{H}$ whereas their assumptions are expressed with the semi-norm used in the definition of the estimator (here d_p). Remark that, with projection semi-norms, the assumptions of Ferraty, Laksaci, and Vieu (2006) and Ferraty, Laksaci, Tadj, et al. (2010) imply that the function F^x only depends on $(\langle x, e_j\rangle)_{1\leq j\leq p}$. Indeed, if we take x and x' such that $\langle x, e_j\rangle = \langle x', e_j\rangle$ for $j = 1, \ldots, p$ (but $\langle x, e_j\rangle \neq \langle x', e_j\rangle$ for a $j > p$), both (H2) and (H2') imply that $F^x(y) = F^{x'}(y)$ for all y. Our assumption is then less restrictive.

Remark 6: Notice that the estimator (3.4) is not consistent when p is fixed. This is also noted by Mas (2012) in a regression setting (see Remark 2, p.4). It is coherent with the fact that we loose information when we project the data. Indeed, suppose that the signal X lies a.s. in $(\text{span}\{e_1, \ldots, e_p\})^\perp$, then $d_p(X_i, x) = \sqrt{\sum_{j=1}^p \langle x, e_j\rangle^2}$ a.s. and $\widehat{F}_{h,p}^x(y) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i\leq y\}}$ if $K_h\left(\sqrt{\sum_{j=1}^p \langle x, e_j\rangle^2}\right) \neq 0$ and 0 otherwise. The bias of such an estimator is then constant and non null as soon as there exists $F^x(y) \neq \mathbb{P}(Y \leq y)$ on a subset of D of positive Lebesgue measure. Hence, if $p < +\infty$, the resulting estimator is not consistent as soon as there exists a $j > p$ such that $\sigma_j > 0$.

3.5.3 Discussion

The rates obtained can be compared to the lower bounds given in Table 3.1 in the Gaussian case under assumptions $H_{X,F}$ and $H_{X,M}$.

3.5.3.1 Comparison with the rates obtained under Assumption $H_{X,F}$

We start from the Karhunen-Loève decomposition of X defined in (3.12). For a Gaussian process, the variables η_j are independent standard normal, $(\lambda_j)_{j\geq 1}$ is a non-increasing sequence of positive numbers and $(\psi_j)_{j\geq 1}$ a basis of $\mathbb{H}$. If $\lambda_{\gamma+1} = 0$ and $\lambda_\gamma > 0$ and if the law of $(\eta_1, \ldots, \eta_\gamma)$ is non-degenerate then Assumption $H_{X,F}$ is fulfilled. Two cases may then occur.

- If $e_j = \pm\psi_j$ for all j, then $\sigma_j^2 = \mathbb{E}[\langle X, e_j\rangle^2] = \mathbb{E}[\langle X, \psi_j\rangle^2] = \lambda_j$ and $\sigma_j = 0$ for $j > \gamma$. Then from Inequality (3.15), with a good choice of (p, h), the integrated risk is upper bounded by $Cn^{-2\beta/(2\beta+\gamma)}$ which fits with the lower bound. According to Inequality (3.14), the pointwise risk is penalized by the term $\sum_{j>p}\langle x_0, e_j\rangle^2$ and the minimax rate is attained only if $x_0 \in \text{span}\{e_1, \ldots, e_p\}$.
- However, if the basis $(e_j)_{j\geq 1}$ is not well-chosen for instance if $e_j = \psi_j$ for $j \notin \{\gamma, \gamma+l\}$ $(l > 0)$, $e_\gamma = \psi_{\gamma+l}$ and $e_{\gamma+l} = \psi_\gamma$, the integrated risk of the estimator is upper bounded by $Cn^{-2\beta/(2\beta+\gamma+l)}$ whereas the minimax rate is $n^{-2\beta/(2\beta+\gamma)}$. It seems that $\widehat{F}_{h,p}$ can not attain the minimax rate in that case.

3.5.3.2 Comparison with rates obtained under Assumption $H_{X,M}$

Thanks to Proposition 2, we are able to obtain the rate of convergence for the estimator.

Corollary 1. *Suppose that the assumptions of Proposition 2 are fulfilled and that $(\xi_1, \ldots, \xi_p)$ admits a density f_p with respect to the Lebesgue measure on $\mathbb{R}^p$ which is continuous in a neighbourhood of 0 and verifies*

$$f_p(0) > 0.$$

Assume also that there exist $\delta > 1$, $c > 0$ such that $\sum_{j>p}\sigma_j^2 \leq cp^{-2\delta+1}$.

(i) Then, for all $x_0 \in \mathbb{H}$ such that there exist $\delta' > 1$, $c' > 0$ such that $\sum_{j>p}\langle x_0, e_j\rangle^2 \leq c'p^{-\delta'+1}$ we have

$$\mathbb{E}\left[\|\widehat{F}_{h,p}^{x_0} - F^{x_0}\|_D^2\right] \leq C\left(\frac{\ln(n)}{\ln(\ln(n))}\right)^{\beta(1-2\min\{\delta,\delta'\})},$$

for a well-chosen bandwidth h and a good choice of p, and where $C > 0$ is a numerical constant.

(ii) We also have

$$\mathbb{E}\left[\|\widehat{F}_{h,p}^{X'} - F^{X'}\|_D^2\right] \leq C\left(\frac{\ln(n)}{\ln(\ln(n))}\right)^{\beta(1-2\delta)},$$

for a well-chosen bandwidth h and a good choice of p, and where $C > 0$ is a numerical constant.

If $cj^{-2a} \leq \lambda_j \leq Cj^{-2a}$, for two constants $c, C > 0$, then Assumption $H_{X,M}$ is fulfilled with $\alpha = 1/(a - 1/2)$, the estimator converges with the minimax rate if $\delta = a$ (adding the condition $\delta' \geq a$ for the pointwise risk). The conclusion is similar to Paragraph 3.5.3.1: if $e_j = \pm\psi_j$ for all $j \geq 1$ (recall that this condition is unrealistic since in most cases the basis $(\psi_j)_{j\geq 1}$ is unknown) then we can choose p and h such that the minimax rate is achieved, up to a logarithmic factor, for the integrated risk and the pointwise risk under an additional condition on x_0. Otherwise, we do not know if the minimax rate can be achieved.

3.6 Numerical study

3.6.1 Simulation of (X, Y)

We simulate X in the following way

$$X(t) := \sum_{j=1}^{J} \sqrt{\lambda_j}\xi_j\psi_j(t) + \xi_0,$$

where $(\lambda_j)_{j\geq 1}$ is a sequence of positive real numbers such that

$$\sum_{j\geq 1} \lambda_j < +\infty,$$

$(\xi_j)_{j\geq 0}$ is a sequence of i.i.d. standard normal random variables and

$$\psi_j(t) := \sqrt{2}\sin(\pi(j - 0.5)t).$$

We can note that the simulated process X is Gaussian and takes values almost surely in a finite-dimensional space. Hence, Assumption $H_{X,F}$ (see section 3.4.1) is verified for $\gamma = J$. However, if J is chosen sufficiently large such that $\sum_{j>J} \lambda_j$ is negligible with respect to $\sum_{j=1}^{J} \lambda_j$ (for instance $J = 150$), we have

$$X(t) = \sum_{j=1}^{J} \sqrt{\lambda_j}\xi_j\psi_j(t) \approx \sum_{j\geq 1} \sqrt{\lambda_j}\xi_j\psi_j(t)$$

and X has a behaviour very similar to the one of an infinite-dimensional process. In particular,

$$\lambda_j = j^{-2},$$

corresponds to a Brownian motion, which verifies Assumption $H_{X,L}$, and

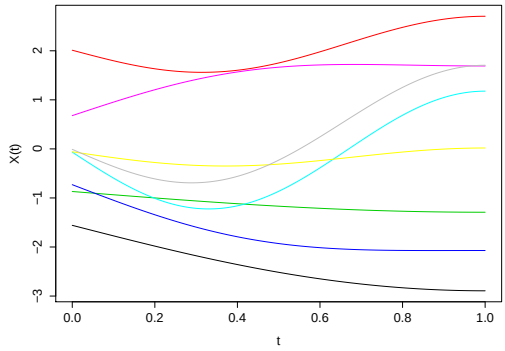
$$\lambda_j = e^{-2j}/j$$

corresponds to a process verifying Assumption $H_{X,M}$. We consider four examples of such processes represented in Figure 3.1.

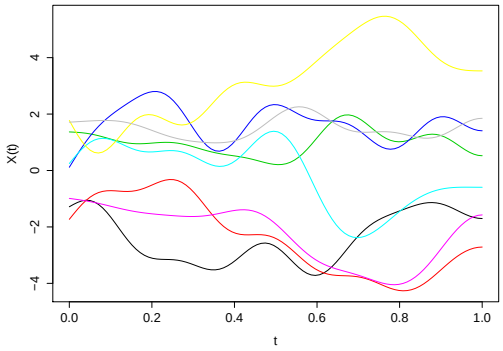
We generate samples following the two models below:

Example 1 (regression model) $Y = \langle\beta, X\rangle^2 + \varepsilon$ with $\beta(t) = \sin(4\pi t)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$,

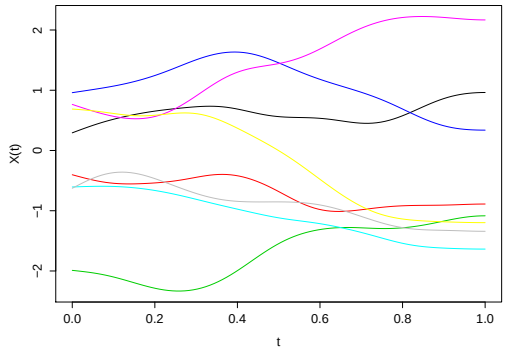
$\lambda_j = j^{-2}$ and $J = 2$ ($H_{X,F}$, $\gamma = 3$)



$\lambda_j = j^{-2}$ and $J = 10$ ($H_{X,F}$, $\gamma = 11$)



$\lambda_j = e^{-2j}/j$ and $J = 150$ ($H_{X,M}$)



$\lambda_j = j^{-2}$ and $J = 150$ ($H_{X,L}$)

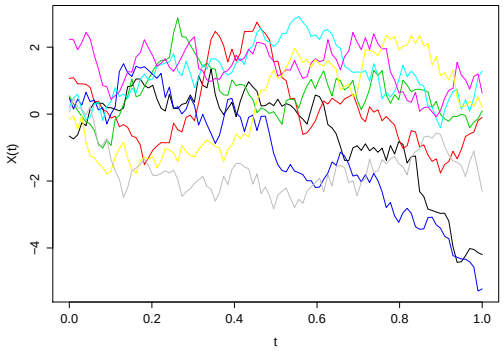


Figure 3.1: Realisations of X.

where ε is independent of X and $\sigma^2 = 0.1$. With this model, the cumulative distribution function of Y given $X = x$ is

$$F^{1,x}(y) := \Phi((y - \langle \beta, x \rangle^2)/\sigma),$$

where $\Phi(z) = \mathbb{P}(Z \leq z)$ for a $Z \sim \mathcal{N}(0,1)$;

Example 2 (Gaussian mixture model) The conditional distribution of Y given $X = x$ is $0.5\mathcal{N}(8 - 4\|x\|, 1) + 0.5\mathcal{N}(8 + 4\|x\|, 1)$. Then the cumulative distribution function of Y given $X = x$ we seek to estimate is

$$F^{2,x}(y) := 0.5\Phi(8 - 4\|x\|) + 0.5\Phi(8 + 4\|x\|).$$

3.6.2 Choice of simulation parameters

Choice of kernel

We choose the uniform kernel (which satisfies Assumption H_K)

$$K(t) = \begin{cases} 1 & \text{if } t \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

Collection of bandwidth

The idea is to define a collection $\mathcal{H}_n$ of the form

$$\mathcal{H}_n := \{C/k, k = 1, \ldots, k_{\max}\}.$$

We see that, with our definition of K, it is unnecessary to consider a bandwidth h which is greater than $\max\{\|X_i - x_0\|, i = 1, \ldots, n\}$. Indeed, for all $h > \max\{\|X_i - x_0\|, i = 1, \ldots, n\}$ we have $K_h(\|X_i - x_0\|) = 1$ and $\widehat{F}_h(y) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i \leq y\}}$. Then we set $C = \max\{\|X_i - x_0\|, i = 1, \ldots, n\}$.

We want to ensure that $\widehat{\varphi}(h)$ is not too small in order to fulfill Assumption H_{b2} and also to avoid instabilities in the calculation of $\widehat{V}(h)$. Then we choose $k_{\max}$ so that, for all $h \in \mathcal{H}_n$,

$$\widehat{\varphi}(h) \geq \frac{\ln n}{n}.$$

We can check that this is verified if $C/k_{\max}$ is the quantile of order $\ln n/n$ of $\{\|X_i\|, i = 1, \ldots, n\}$, this is the choice we make in the following.

Calibration of κ

The parameter κ appearing in the bandwidth selection criterion influences the quality of estimation. Indeed, recall that

$$\widehat{h} \in \arg\min_{h \in \mathcal{H}_n} \left\{ A(h) + \kappa \frac{\ln n}{n\widehat{\varphi}(h)} \right\},$$

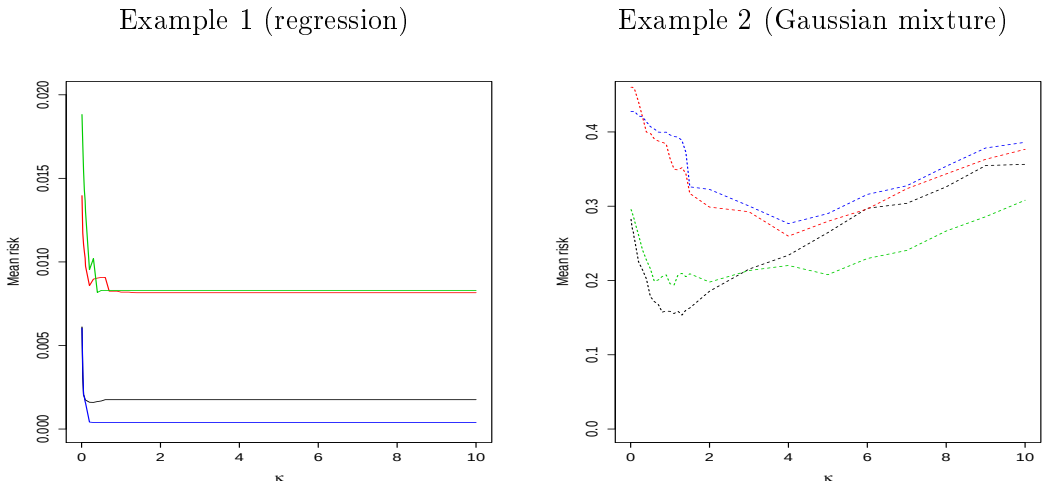


Figure 3.2: Plot of approximated mean risk with respect to κ. Black curves: $H_{X,F}$ with $\gamma = 2$, red curves: $H_{X,F}$ with $\gamma = 10$, green curves: $H_{X,M}$ and blue curves: $H_{X,L}$, $n = 500$.

where $A(h)$ estimates the bias of $\widehat{F}_h$ and $n\widehat{\varphi}(h)$ is the order of the variance of $\widehat{F}_h$. Roughly speaking:

- If κ is small, then $A(h)$ is the most influential term. Since this term tends to decrease with h, we select small bandwidths;
- If κ is large, the reverse occurs: we select large bandwidths.

We perform a Monte-Carlo study for different values of κ. For each set of parameters and each value of κ, the risk $\mathbb{E}\left[\|\widehat{F}^{X'}_{\widehat{h}} - F^{X'}\|^2_D\right]$ is approximated by the mean of $N = 50$ Monte-Carlo replications of the random variable $\|\widehat{F}^{X'}_{\widehat{h}} - F^{X'}\|^2_D$. We fix $D := [\min_{i=1,\dots,n}\{Y_i\}, \max_{i=1,\dots,n}\{Y_i\}]$ and B is chosen sufficiently large in order to ensure that $X' \in B$ in all the simulation study. In the light of the results represented in Figure 3.2, we fix $\kappa = 4$. However, we see here that the calibration of the constant κ is a real problem. Data-driven calibration tools developed in model selection contexts, such as slope heuristic (see section 2.4.3, p.42), could be very useful here. However, this kind of tool has not been developed yet in our bandwidth selection context.

3.6.3 Results

We see on Figure 3.3 that the estimation is quite stable for the regression model (example 1) and that the estimator we select is not far (and sometimes is identical) to the oracle (Figure 3.4). However we see in figures 3.5 and 3.6 that the estimation is harder for the Gaussian mixture model in the cases $H_{X,F}$ with $\gamma = 10$ and $H_{X,L}$, since our criterion select models with a too small h. Note that our bandwidth selection criterion behaves better in the cases $H_{X,F}$ and $H_{X,M}$.

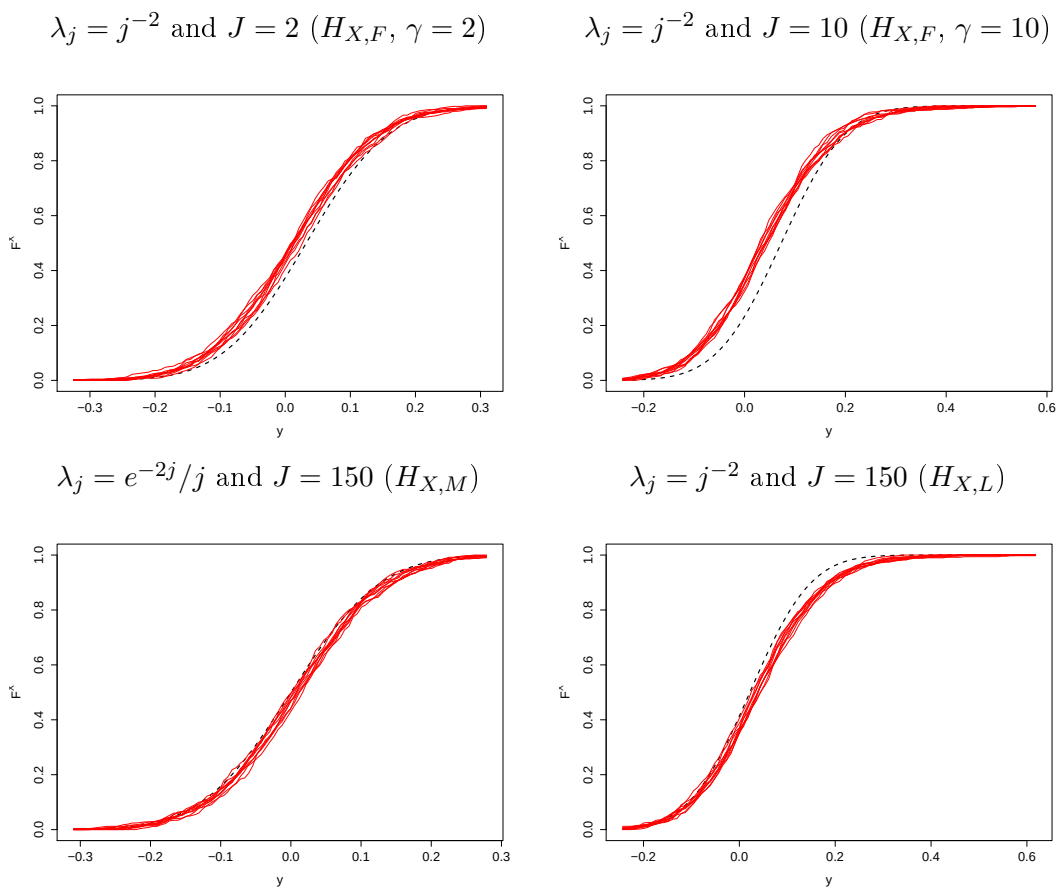


Figure 3.3: Plot of $\widehat{F}^{1,x_0}_{\widehat{h}}$ calculated from 10 independent samples (red curves) where x_0 is a copy of X, $n = 500$. The dotted black curve represents F^{1,x_0}.

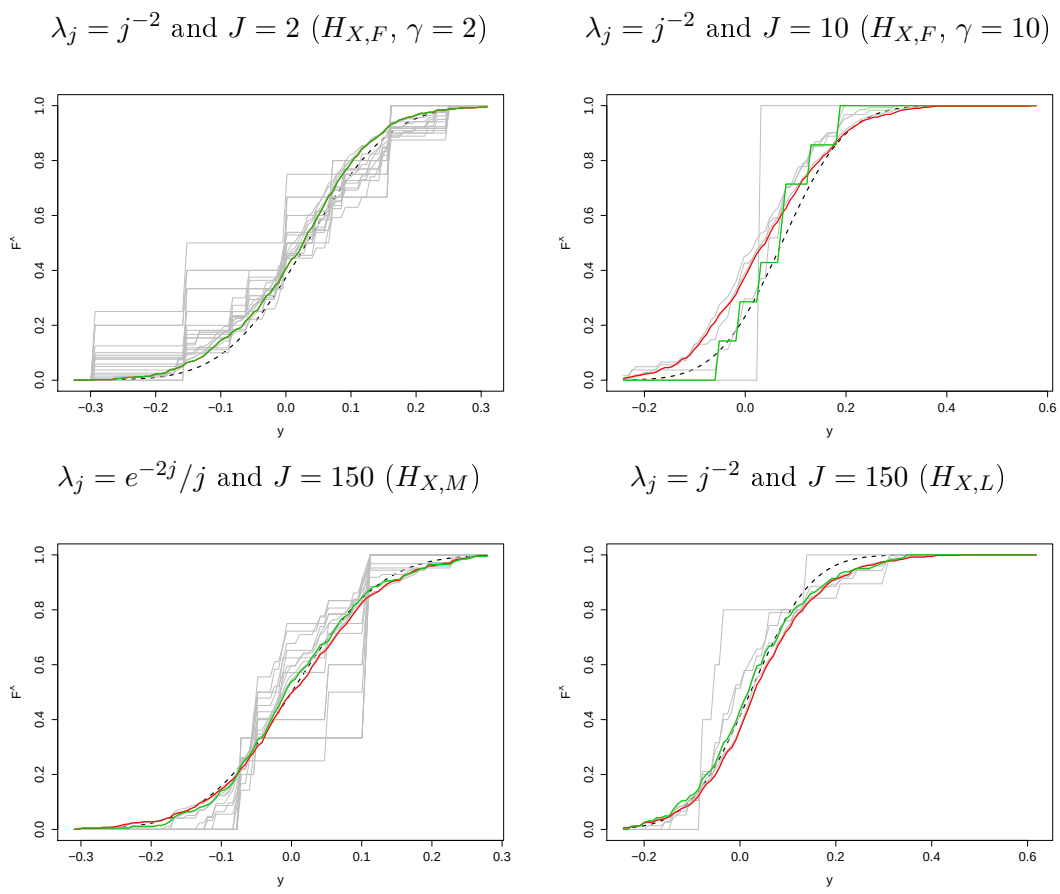


Figure 3.4: Plot of $\widehat{F}_h^{1,x_0}$ for all $h \in \mathcal{H}_n$ (gray curves), $n = 500$. The red curve represents $\widehat{F}_{\widehat{h}}^{1,x_0}$ and the green curve is the pseudo-oracle $\widehat{F}_{h^*}^{1,x_0}$ where $h^* := \arg\min_{h \in \mathcal{H}_n} \left\{ \left\| \widehat{F}_h^{1,x_0} - F^{1,x_0} \right\|_D^2 \right\}$.

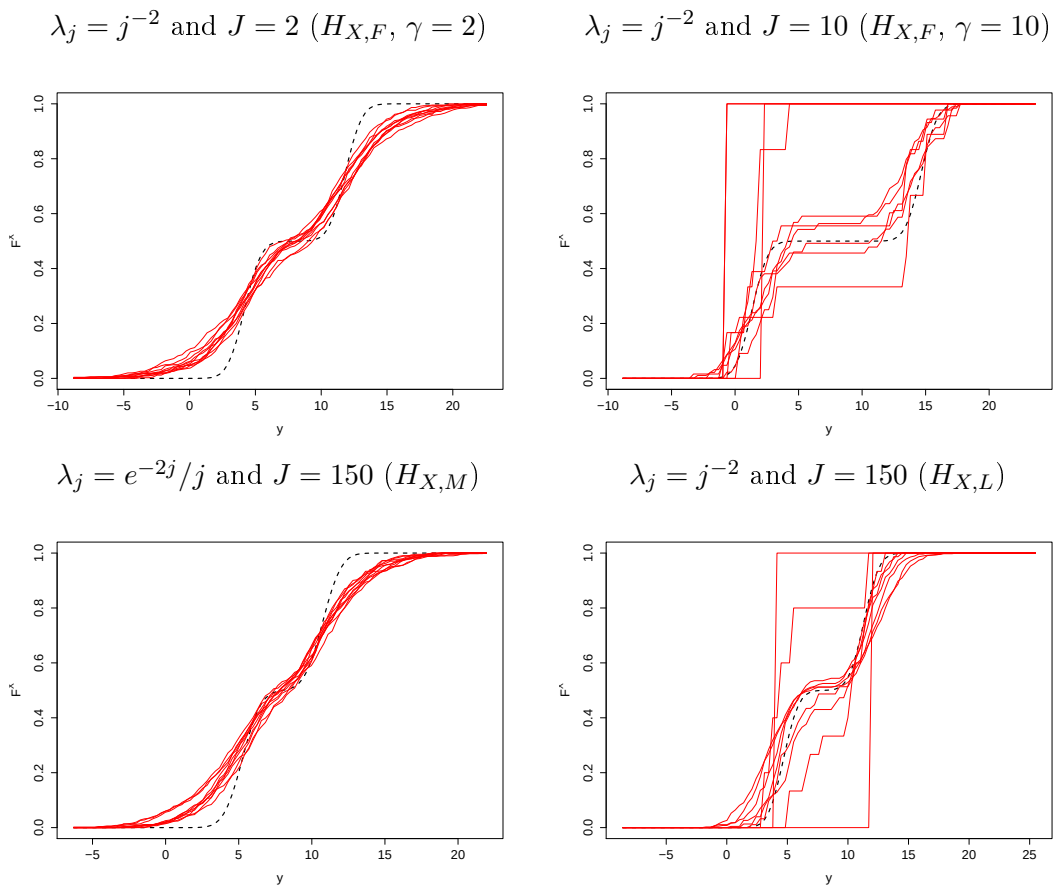


Figure 3.5: Plot of $\widehat{F}^{2,x_0}_{\widehat{h}}$ calculated from 10 independent samples (red curves) where x_0 is a copy of X, $n = 500$. The dotted black curve represents F^{2,x_0}.

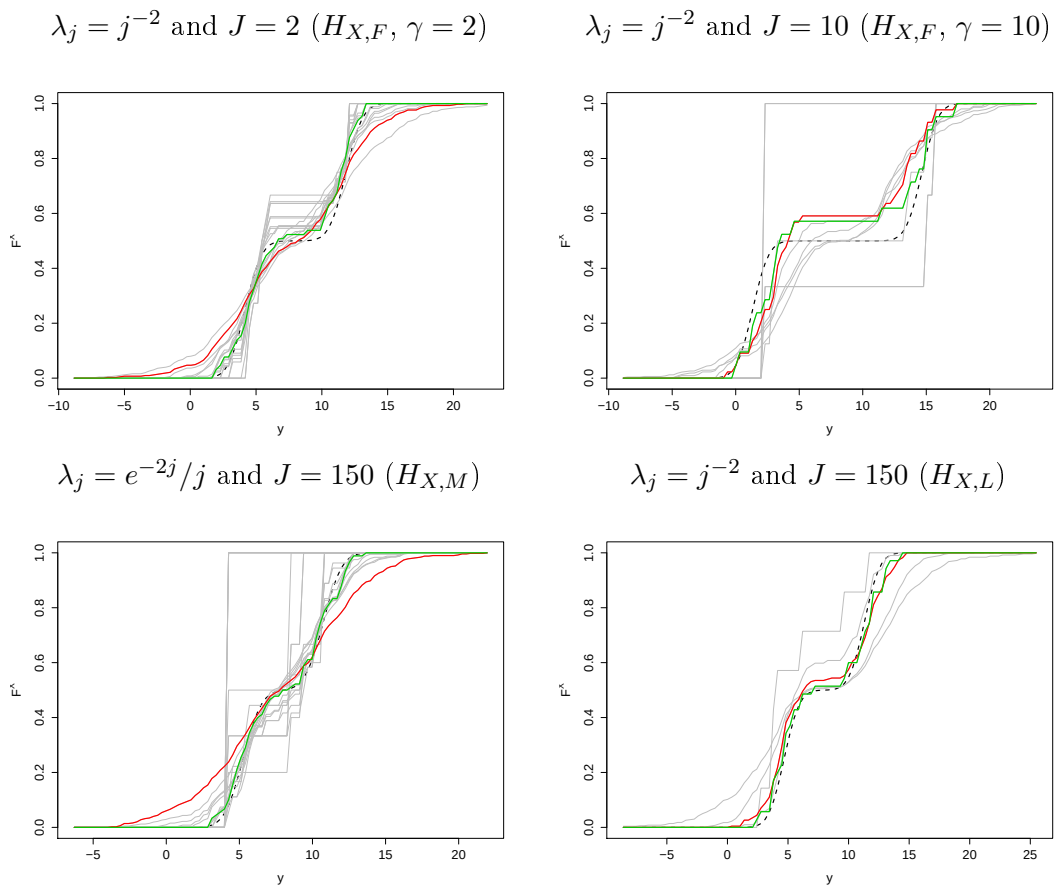


Figure 3.6: Plot of $\widehat{F}_h^{2,x_0}$ for all $h \in \mathcal{H}_n$ (gray curves), $n = 500$. The red curve represents $\widehat{F}_{2,\widehat{h}}^{x_0}$ and the green curve is the pseudo-oracle $\widehat{F}_{h^*}^{2,x_0}$ where $h^* := \arg\min_{h \in \mathcal{H}_n} \left\{ \left\| \widehat{F}_h^{2,x_0} - F^{2,x_0} \right\|_D^2 \right\}$.

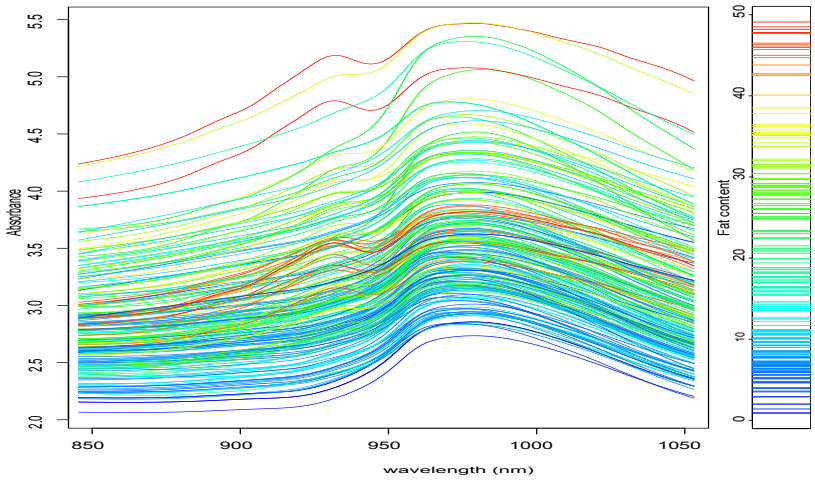


Figure 3.7: Curves of the sample (left). The colors of each curve is defined by its corresponding fat content (right)

3.6.4 Application to spectrometric dataset

The data we study, available online [3], are recorded on Tecator Infratec Food and Feed Analyzer. For each unit i ($i = 1, \ldots, n$, with $n = 215$), we observe one spectrometric curve X_i which corresponds to the absorbances of a piece of chopped meat measured at 100 wavelengths ranging between 850 and 1050 nm. The aim is to study the link between a spectrometric curve of a chopped meat and its fat content. For each curve X_i, we denote by Y_i the corresponding fat content.

These data have been widely studied, mainly in regression contexts (see Ferraty and Vieu (2002, 2006) and Ferraty, Mas, and Vieu (2007)).

In a first step, the data are centred. Then we apply our estimation procedure. We choose ten curves $x_0^{(j)} := X(i_0^{(j)})$ ($j \in \{1, \ldots, 10\}$) randomly in the sample. For each curve $x_0^{(j)}$ we estimate the conditional distribution of Y given $X = x_0^{(j)}$ using the information of $\{(X_i, Y_i), i \neq i_0^{(j)}\}$. The results are given in Figure 3.8. We see that when $Y_{i_0^{(j)}}$ is small (blue curves), the estimated conditional distribution function $y \mapsto \widehat{F}_{\widehat{h}}^{x_0^{(j)}}$ is strongly increasing in the interval $[0, 10]$, which indicates that our estimator detects that Y must be small with large probability. Conversely, when $Y_{i_0^{(j)}}$ is large (red curves) the estimated conditional distribution function faster increases in the interval [30,50]. These results indicate that our estimators are able to capture the repartition of Y given $X = x_0$ when x_0 is taken in the sample.

3. http://www.math.univ-toulouse.fr/~ferraty/SOFTWARES/NPFDA/

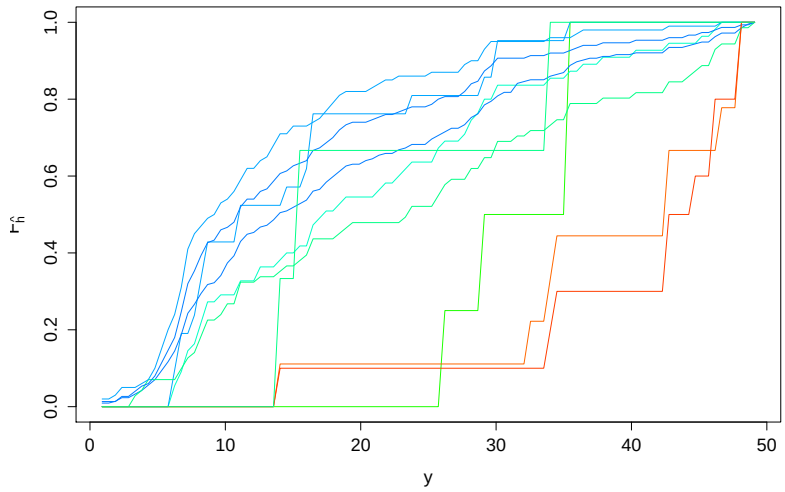


Figure 3.8: Plot of $\widehat{F}^{x_0^{(j)}}_{\widehat{h}}$, for $j = 1, \ldots, 10$. The color of each plot is defined by the fat content $Y_{i_0^{(j)}}$ as represented in Figure 3.7 – right.

3.7 Proofs

We will mainly focus on the proof of the results for the integrated risk (since it is the one for which adaptation results are provided), and only highlight the differences when choosing the pointwise criterion.

We denote by $\mathbb{E}_{X'}$ (resp. $\mathbb{P}_{X'}$, $\mathrm{Var}_{X'}$) the conditional expectation (resp. probability, variance) given X'. We also introduced the classical norm $\|.\|_{L^q(\mathbb{R})}$ of the space $L^q(\mathbb{R})$ of integrable functions (the notation will be used with $q = 2$ and $q = \infty$).

Recall that $K_h(x) := h^{-1}K(h^{-1}x)$. Assumptions H_K and H_φ imply that, for all $l \geq 1$,

$$h^{-l} m_l \varphi(h) \mathbf{1}_B(X') \leq \mathbb{E}_{X'}\left[K_h^l(d(X, X'))\right] \mathbf{1}_B(X') \leq h^{-l} M_l \varphi(h) \mathbf{1}_B(X') \text{ a.s.} \tag{3.16}$$

where $m_l := c_K^l c_\varphi$ and $M_l := C_K^l C_\varphi$. These inequalities are useful in the sequel.

3.7.1 A preliminary result

One of the key arguments in the proofs of Theorems 3, 4, and Proposition 2 is the control of the deviations (in probability and expectation) of the process R_h^x, for $x \in \mathbb{H}$, defined by

$$R_h^x = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \dfrac{K_h(d(X_i, X'))}{\mathbb{E}_{X'}\left[K_h(d(X, X'))\right]}, & \text{if } x = X', \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \dfrac{K_h(d(X_i, x))}{\mathbb{E}\left[K_h(d(X, x))\right]}, & \text{if } x \in \mathbb{H} \text{ is fixed.} \end{cases} \tag{3.17}$$

The following lemma establishes the result which is useful to control the integrated risk of the estimators. The proof can be found below.

Lemma 8. *Assume H_K and H_φ. For any $\eta > 0$, on the set $\{X' \in B\}$, the following inequality holds a.s.*

$$\mathbb{P}_{X'}\left(\left|R_h^{X'} - 1\right| > \eta\right) \leq 2\exp\left(-\frac{n\eta^2\varphi(h)}{2\left(\frac{M_2}{m_1^2} + \frac{C_K\eta}{m_1}\right)}\right). \tag{3.18}$$

Moreover, assume also H_{b_2}, and denote by $V_R(h) = \kappa_R \ln(n)/(n\varphi(h))$, we have a.s.

$$\mathbb{E}_{X'}\left[\left(\left(R_h^{X'} - 1\right)^2 - V_R(h)\right)_+\right] \leq \min\left(\frac{4M_2}{m_1^2}, \frac{64C_K^2}{m_1^2}\right)\frac{1}{n^\alpha}, \tag{3.19}$$

for any $\alpha > 0$, as soon as $\kappa_R > \max(4M_2\alpha/m_1^2, 32C_K^2\alpha^2/m_1^2C_0)$.

Fix a point $x_0 \in \mathbb{H}$. Then Inequality (3.18) becomes

$$\mathbb{P}\left(\left|R_h^{x_0} - 1\right| > \eta\right) \leq 2\exp\left(-\frac{n\eta^2\varphi^{x_0}(h)}{2\left(\frac{C_K^2}{c_K^2} + \frac{C_K\eta}{c_K}\right)}\right). \tag{3.20}$$

3.7.1.1 Proof of Lemma 8

To prove Inequality (3.18), the guideline is to apply Bernstein's Inequality (see Lemma 22 p. 180), for the conditional probability $\mathbb{P}_{X'}$, with $T_i = K_h(d(X_i, X'))/\mathbb{E}_{X'}[K_h(d(X_i, X'))]$, and $R_h^{X'} - 1 = S_n(T)/n$ (recall that we consider here conditional expectation and probability with respect to X'). Let us compute the quantities v and b_0 involved in the inequality. First, on the set $\{X' \in B\}$, Inequality (3.16) implies that

$$\mathrm{Var}_{X'}(T_1) \leq \mathbb{E}_{X'}\left[T_1^2\right] = \frac{\mathbb{E}_{X'}\left[K_h^2(d(X_1, X'))\right]}{\left(\mathbb{E}_{X'}\left[K_h(d(X_1, X'))\right]\right)^2} \leq \frac{h^{-2}M_2\varphi(h)}{\left(h^{-1}m_1\varphi(h)\right)^2} = \frac{M_2}{m_1^2}\frac{1}{\varphi(h)} =: v^2.$$

Similarly, for $l \geq 2$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{X'}\left[|T_i|^l\right] &= \mathbb{E}_{X'}\left[|T_1|^l\right] = \frac{\mathbb{E}_{X'}\left[K_h^l(d(X_1, X'))\right]}{\left(\mathbb{E}_{X'}\left[K_h(d(X_1, X'))\right]\right)^l} \\ &\leq \frac{h^l M_l\varphi(h)}{\left(h\varphi(h)m_1\right)^l} = \frac{M_l}{m_1^l}\frac{1}{\varphi^{l-1}(h)}.\end{aligned}$$

By splitting $M_l = C_K^l C_\varphi = M_2 C_K^{l-2}$, the last upper bound can be written

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{X'}\left[|T_i|^l\right] \leq \frac{M_2}{m_1^2}\frac{1}{\varphi(h)}\frac{C_K^{l-2}}{m_1^{l-2}}\frac{1}{(\varphi(h))^{l-2}} = v^2 b_0^{l-2},$$

with $b_0 = C_K/(m_1\varphi(h))$. We now apply the first inequality of Lemma 22, this complete the proof of Inequality (3.18). The proof may be adapted easily to demonstrate Inequality (3.20). For Inequality (3.19), we follow the same strategy as Comte and Genon-Catalot (2012), pages

20-21. First

$$\begin{aligned}
\mathbb{E}_{X'}\left[\left(\left(R_h^{X'}-1\right)^2 - V_R(h)\right)_+\right] &= \int_0^\infty \mathbb{P}_{X'}\left(\left(\left(R_h^{X'}-1\right)^2 - V_R(h)\right)_+ \geq u\right) du \\
&\leq \int_0^\infty \mathbb{P}_{X'}\left(\left|R_h^{X'}-1\right| \geq \sqrt{V_R(h)+u}\right) du \\
&\leq 2\min\left\{\int_0^\infty \exp\left(-\frac{n(u+V_R(h))}{4v^2}\right) du, \right. \\
&\qquad \left. \int_0^\infty \exp\left(-\frac{n\sqrt{u+V_R(h)}}{4b_0}\right) du\right\},
\end{aligned}$$

thanks to Inequality (B.1). Now,

$$\frac{n(u+V_R(h))}{4v^2} = n\varphi(h)u\frac{m_1^2}{4M_2} + \ln(n)\frac{m_1^2\kappa_R}{4M_2},$$

which leads to

$$\begin{aligned}
\int_0^\infty \exp\left(-\frac{n(u+V_R(h))}{4v^2}\right) du &= n^{-\frac{m_1^2\kappa_R}{4M_2}} \int_0^\infty \exp\left(-n\varphi(h)\frac{m_1^2}{4M_2}u\right) du, \\
&\leq \frac{4M_2}{m_1^2}\frac{1}{n^{1+\kappa_R m_1^2/4M_2}}\frac{1}{\varphi(h)}.
\end{aligned}$$

Since, by Assumption H_{b_2}, $\varphi(h) \geq C_0\ln(n)/n$, we obtain

$$\int_0^\infty \exp\left(-\frac{n(u+V_R(h))}{4v^2}\right) du \leq \frac{4M_2}{C_0 m_1^2}\frac{1}{\ln(n)n^{\kappa_R m_1^2/4M_2}},$$

and the last upper bound is smaller than $(4M_2/C_0m_1^2)/n^\alpha$ as soon as $\kappa_R > 4M_2\alpha/m_1^2$. For the other integral, we begin with a lower bound for $n\sqrt{u+V_R(h)}/4b_0$,

$$\begin{aligned}
\frac{n\sqrt{u+V_R(h)}}{4b_0} &\geq \frac{m_1}{4C_K}n\varphi(h)\frac{1}{\sqrt{2}}\left(\sqrt{V_R(h)}+\sqrt{u}\right), \\
&= \frac{m_1}{4\sqrt{2}C_K}\sqrt{\kappa_R}\sqrt{\ln(n)}\sqrt{n\varphi(h)} + \frac{m_1}{4\sqrt{2}C_K}n\varphi(h)\sqrt{u}, \\
&\geq \frac{m_1\sqrt{C_0}}{4\sqrt{2}C_K}\sqrt{\kappa_R}\ln(n) + \frac{m_1 C_0}{4\sqrt{2}C_K}\ln(n)\sqrt{u},
\end{aligned}$$

by using $\varphi(h) \geq C_0 \ln(n)/n$ another time. Thus,

$$\begin{aligned}\int_0^\infty \exp\left(-\frac{n\sqrt{u+V_R(h)}}{4b_0}\right) du &\leq n^{-\frac{m_1\sqrt{C_0}\sqrt{\kappa_R}}{4\sqrt{2}C_K}} \int_0^\infty \exp\left(-\frac{m_1 C_0 \ln(n)}{4\sqrt{2}C_K}\sqrt{u}\right) du \\ &= \frac{64C_K^2}{m_1^2 C_0^2}\int_0^\infty s\exp(-s)ds \frac{1}{\ln^2(n) n^{\frac{m_1\sqrt{C_0}\sqrt{\kappa_R}}{4\sqrt{2}C_K}}} \\ &= \frac{64C_K^2}{m_1^2 C_0^2}\frac{1}{\ln^2(n) n^{\frac{m_1\sqrt{C_0}\sqrt{\kappa_R}}{4\sqrt{2}C_K}}} \leq \frac{64C_K^2}{m_1^2 C_0^2}\frac{1}{n^\alpha},\end{aligned}$$

as soon as $\kappa_R > 32C_K^2\alpha^2/m_1^2 C_0$. This ends the proof of Lemma 8.

3.7.2 Proof of Theorem 3

3.7.2.1 Main part of the proof of the Inequality (3.7)

Following Ferraty, Laksaci, and Vieu (2006) and Ferraty, Laksaci, Tadj, et al. (2010), we define

$$\widetilde{F}_h^{X'}(y) := \sum_{i=1}^n \widetilde{W}_h^{(i)}(X')\mathbf{1}_{\{Y_i\leq y\}}, \text{ where } \widetilde{W}_h^{(i)}(X') = \frac{K_h\left(d(X_i,X')\right)}{n\mathbb{E}_{X'}\left[K_h\left(d(X_1,X')\right)\right]}. \quad (3.21)$$

We also have $R_h^{X'} := \sum_{i=1}^n \widetilde{W}_h^{(i)}(X')$ (see Definition (3.17)). First, notice that since $\widehat{F}_h^{X'} \leq 1$ and $F^{X'} \leq 1$ a.s.,

$$\begin{aligned}\mathbb{E}\left[\|\widehat{F}_h^{X'} - F^{X'}\|_D^2 \mathbf{1}_{\{R_h^{X'}<1/2\}}\mathbf{1}_B(X')\right] &\leq 2\mathbb{E}\left[\left(\|\widehat{F}_h^{X'}\|_D^2 + \|F^{X'}\|_D^2\right)\mathbf{1}_{\{R_h^{X'}<1/2\}}\mathbf{1}_B(X')\right], \\ &\leq 4|D|\mathbb{P}\left(\left\{R_h^{X'} < 1/2\right\} \cap \left\{X' \in B\right\}\right).\end{aligned}$$

Now, with $\mathbb{P}(\{R_h^{X'} < 1/2\} \cap \{X' \in B\}) \leq \mathbb{P}(\{|R_h^{X'} - 1| > 1/2\} \cap \{X' \in B\})$ and with Lemma 8 we get

$$\mathbb{E}\left[\|\widehat{F}_h^{X'} - F^{X'}\|_D^2 \mathbf{1}_{\{R_h^{X'}<1/2\}}\mathbf{1}_B(X')\right] \leq 8|D|\exp\left(-\frac{m_1}{8\left(M_2/m_1 + C_K/2\right)}n\varphi(h)\right) \leq \frac{C}{n\varphi(h)},$$

where $C = 64|D|e^{-1}\frac{M_2/m_1+C_K/2}{m_1}$. The last inequality comes from the bound $xe^{-x} \leq e^{-1}$, $x > 0$.
We must now control $\mathbb{E}\left[\|\widehat{F}_h^{X'} - F^{X'}\|_D^2 \mathbf{1}_B(X')\mathbf{1}_{\{R_h^{X'}\geq 1/2\}}\right]$. Recall that $\widehat{F}_h^{X'} = \widetilde{F}_h^{X'}/R_h^{X'}$. We thus have,

$$\begin{aligned}\mathbb{E}\left[\|\widehat{F}_h^{X'} - F^{X'}\|_D^2 \mathbf{1}_B(X')\mathbf{1}_{\{R_h^{X'}\geq 1/2\}}\right] &\leq 12\mathbb{E}\left[\left\|\widetilde{F}_h^{X'} - \mathbb{E}_{X'}\left[\widetilde{F}_h^{X'}\right]\right\|_D^2 \mathbf{1}_B(X')\right] \\ &+12\mathbb{E}\left[\left\|\mathbb{E}_{X'}\left[\widetilde{F}_h^{X'}\right] - F^{X'}\right\|_D^2\right] \\ &+12\mathbb{E}\left[\left(1-R_h^{X'}\right)^2\left\|F^{X'}\right\|_D^2 \mathbf{1}_B(X')\right]. \quad (3.22)\end{aligned}$$

The first and third terms are variance terms, bounded by Lemmas 9 and 10 proved below. The second one is a bias term, controlled by Lemma 11.

Lemma 9. *Under Assumptions H_K and H_φ, on the set $\{X' \in B\}$,*

$$\mathbb{E}_{X'}\left[\left\|\widetilde{F}_h^{X'} - \mathbb{E}_{X'}\left[\widetilde{F}_h^{X'}\right]\right\|_D^2\right] \leq |D|\frac{M_2}{m_1^2}\frac{1}{n\varphi(h)}.$$

Lemma 10. *Under Assumptions H_K and H_φ,*

$$\mathbb{E}\left[\left(R_h^{X'} - 1\right)^2 \mathbf{1}_B(X')\right] \leq \frac{M_2}{m_1^2}\frac{1}{n\varphi(h)}.$$

Lemma 11. *Under Assumption H_F,*

$$\mathbb{E}\left[\left\|F^{X'} - \mathbb{E}_{X'}\left[\widetilde{F}_h^{X'}\right]\right\|_D^2\right] \leq C_D^2 h^{2\beta}.$$

This ends the proof of Inequality (3.7). The scheme can easily be adapted to prove (3.6).

□

3.7.2.2 Proof of Lemmas 9 and 10 (upper bounds for the variance terms)

Proof of Lemma 9 By Fubini's Theorem

$$\begin{aligned}\mathbb{E}_{X'}\left[\left\|\widetilde{F}_h^{X'} - \mathbb{E}_{X'}\left[\widetilde{F}_h^{X'}\right]\right\|_D^2\right] &= \int_D \mathbb{E}_{X'}\left[\left(\widetilde{F}_h^{X'}(y) - \mathbb{E}_{X'}\left[\widetilde{F}_h^{X'}(y)\right]\right)^2\right] dy \\ &= \int_D \mathrm{Var}_{X'}\left(\widetilde{F}_h^{X'}(y)\right) dy.\end{aligned}$$

Since, for all $y \in D$, $\widetilde{F}_h^{X'}(y)$ is a mean of independent and identically distributed random variables (conditionally to X'), we have, on the set $\{X' \in B\}$,

$$\begin{aligned}\mathbb{E}_{X'}\left[\left\|\widetilde{F}_h^{X'} - \mathbb{E}_{X'}\left[\widetilde{F}_h^{X'}\right]\right\|_D^2\right] &= \frac{1}{n}\int_D \mathrm{Var}_{X'}\left(\frac{K_h\left(d(X_1, X')\right)\mathbf{1}_{\{Y_1 \leq y\}}}{\mathbb{E}_{X'}\left[K_h\left(d(X_1, X')\right)\right]}\right) dy \\ &\leq \frac{|D|}{n}\mathbb{E}_{X'}\left[\frac{K_h^2\left(d(X_1, X')\right)}{\left(\mathbb{E}_{X'}\left[K_h\left(d(X_1, X')\right)\right]\right)^2}\right] \leq \frac{|D|}{n}\frac{M_2}{m_1^2}\frac{1}{\varphi(h)},\end{aligned}$$

where the last inequality comes from Inequality (3.16).

□

Proof of Lemma 10 Since $\mathbb{E}_{X'}\left[R_h^{X'}\right] = 1$, remark that,

$$\begin{aligned}\mathbb{E}\left[\left(R_h^{X'} - 1\right)^2 \mathbf{1}_B(X')\right] &= \mathbb{E}\left[\mathrm{Var}_{X'}\left(R_h^{X'}\right)\mathbf{1}_B(X')\right] \\ &= \frac{1}{n}\mathbb{E}\left[\mathrm{Var}_{X'}\left(\frac{K_h(d(X_1, X'))}{\mathbb{E}_{X'}\left[K_h(d(X_1, X'))\right]}\right)\mathbf{1}_B(X')\right],\end{aligned}$$

and the result comes also from Inequality (3.16).

3.7.2.3 Proof of Lemma 11 (upper bound for the bias term)

First remark that, for $y \in D$, a.s.

$$\mathbb{E}_{X'}\left[\widetilde{F}_h^{X'}(y)\right] = n\mathbb{E}_{X'}\left[\mathbb{E}\left[\widetilde{W}_h^{(1)}(X')\mathbf{1}_{\{Y_1\leq y\}}|X_1\right]\right] = n\mathbb{E}_{X'}\left[\widetilde{W}_h^{(1)}(X')F^{X_1}(y)\right]$$

and since $n\mathbb{E}_{X'}\left[\widetilde{W}_h^{(1)}(X')F^{X'}\right] = F^{X'}$,

$$F^{X'}(y) - \mathbb{E}_{X'}\left[\widetilde{F}_h^{X'}(y)\right] = n\mathbb{E}_{X'}\left[\widetilde{W}_h^{(1)}(X')\left(F^{X'}(y) - F^{X_1}(y)\right)\right].$$

Then,

$$\begin{aligned}\mathbb{E}\left[\left\|F^{X'} - \mathbb{E}_{X'}\left[\widetilde{F}_h^{X'}\right]\right\|_D^2\right] &\leq n^2\mathbb{E}\left[\mathbb{E}_{X'}\left[\widetilde{W}_h^{(1)}(X')\left\|F^{X_1} - F^{X'}\right\|_D\right]^2\right] \\ &\leq C_D^2 n^2\mathbb{E}\left[\mathbb{E}_{X'}\left[\widetilde{W}_h^{(1)}(X')\left\|X_1 - X'\right\|^\beta\right]^2\right], \qquad (3.23)\end{aligned}$$

by H_F. Now, since K is supported on $[0,1]$, if $d^2(X_1,X') = \|X_1 - X'\|^2 > h$ then $\widetilde{W}_h^{(1)}(X') = 0$,

$$\mathbb{E}\left[\left\|F^{X'} - \mathbb{E}_{X'}\left[\widetilde{F}_h^{X'}\right]\right\|_D^2\right] \leq C_D^2 n^2\mathbb{E}\left[\mathbb{E}_{X'}\left[\widetilde{W}_h^{(1)}(X')h^\beta\right]^2\right].$$

But $\mathbb{E}_{X'}\left[\widetilde{W}_h^{(1)}(X')\right] = 1/n$, which ends the proof:

$$\mathbb{E}\left[\left\|F^{X'} - \mathbb{E}_{X'}\left[\widetilde{F}_h^{X'}\right]\right\|_D^2\right] \leq C_D^2 h^{2\beta}.$$

.

□

3.7.3 Proof of an intermediate result for Theorem 4: the case of known small ball probability

This section is an introduction to the proof of Theorem 4: we first deal with the toy case of known small ball probability. It is thus possible to define a selection rule by $\tilde{h} = \text{argmin}_{h\in\mathcal{H}_n}\{A(h) + V(h)\}$, with

$$V(h) = \kappa\frac{\ln(n)}{n\varphi(h)} \text{ and } A(h) = \max_{h'\in\mathcal{H}_n}\left(\left\|\widehat{F}_{h'}^{X'} - \widehat{F}_{h\vee h'}^{X'}\right\|_D^2 - V(h')\right)_+. \qquad (3.24)$$

Compared to the data-driven criterion (3.10), the variance term $V(h)$ is deterministic here.

Assume that H_K, H_φ, H_F, H_{b1} and H_{b2} hold. The pseudo-estimator $\widehat{F}_{\tilde{h}}$ is such that

$$\mathbb{E}\left[\left\|\widehat{F}_{\tilde{h}}^{X'} - F^{X'}\right\|_D^2 \mathbf{1}_B(X')\right] \leq c \min_{h\in\mathcal{H}_n}\left\{h^{2\beta} + \frac{\ln(n)}{n\varphi(h)}\right\} + \frac{C}{n}, \tag{3.25}$$

where c and C are constants which depend on c_K, C_K, c_φ, C_φ, $|D|$, and C_D.

The proof of such inequality is simpler than Theorem 4, and is a good illustration of the model selection tools required to deal with a data-driven selected bandwidth. To prove Theorem 4, we will then come down to Inequality (3.25).

Main part of the proof of Inequality (3.25)

Let $h \in \mathcal{H}_n$ be fixed. We start with the following decomposition for the loss of the estimator$\widehat{F}_{\tilde{h}}^{X'}$:

$$\left\|\widehat{F}_{\tilde{h}}^{X'} - F^{X'}\right\|_D^2 \leq 3\left\|\widehat{F}_{\tilde{h}}^{X'} - \widehat{F}_{\tilde{h}\vee h}^{X'}\right\|_D^2 + 3\left\|\widehat{F}_{\tilde{h}\vee h}^{X'} - \widehat{F}_h^{X'}\right\|_D^2 + 3\left\|\widehat{F}_h^{X'} - F^{X'}\right\|_D^2.$$

The definitions of $A(h)$, $A(\tilde{h})$ and then the one of $\tilde{h}$ enable to write

$$\begin{aligned} 3\left\|\widehat{F}_{\tilde{h}}^{X'} - \widehat{F}_{\tilde{h}\vee h}^{X'}\right\|_D^2 + 3\left\|\widehat{F}_{\tilde{h}\vee h}^{X'} - \widehat{F}_h^{X'}\right\|_D^2 &\leq 3\left(A(h) + V\left(\tilde{h}\right)\right) + 3\left(A\left(\tilde{h}\right) + V(h)\right) \\ &\leq 6\left(A(h) + V(h)\right). \end{aligned}$$

Besides, the quantity $\|\widehat{F}_h^{X'} - F^{X'}\|_D^2$ is the loss of an estimator with fixed bandwidth h and has already been bounded (see Theorem 3 Inequality (3.7)). Hence we obtain

$$\mathbb{E}\left[\left\|\widehat{F}_{\tilde{h}}^{X'} - F^{X'}\right\|_D^2 \mathbf{1}_B(X')\right] \leq 6\mathbb{E}\left[A(h)\mathbf{1}_B(X')\right] + 6V(h) + 3C\left(h^{2\beta} + \frac{1}{n\varphi(h)}\right), \tag{3.26}$$

where C is the constant of Theorem 3 (Inequality (3.7)). The remainding part of the proof is the result of the lemma hereafter, the proof of which is postponed to the following section.

Lemma 12. *Let $h \in \mathcal{H}_n$ be fixed. Under the assumptions of Theorem 4, there exist two constants C and C_1 such that,*

$$\mathbb{E}\left[A(h)\mathbf{1}_B(X')\right] \leq C_1 h^{2\beta} + \frac{C_2}{n}. \tag{3.27}$$

The constant C_2 depends on C_0, $|D|$, M_2, m_1 and C_K and the constant C_1 only depends on C_D.

Applying Inequality (3.27) in (3.26) implies Inequality (3.25) by taking the infimum over $h \in \mathcal{H}_n$.

□

Proof of Lemma 12 (Upper bound for $A(h)$)

Fix $h, h' \in \mathcal{H}_n$. We define the set $\Omega_{h,h'} = \{R_{h'}^{X'} \geq 1/2\} \cap \{R_{h\vee h'}^{X'} \geq 1/2\}$ and split

$$\left\|\widehat{F}_{h'}^{X'} - \widehat{F}_{h\vee h'}^{X'}\right\|_D^2 \leq \left\|\widehat{F}_{h'}^{X'} - \widehat{F}_{h\vee h'}^{X'}\right\|_D^2 \left(\mathbf{1}_{\Omega_{h,h'}} + \mathbf{1}_{\Omega_{h,h'}^c}\right).$$

Recall that we write the estimator $\widehat{F}_h^{X'}(y) = \widetilde{F}_h^{X'}(y)/R_h^{X'}$, with $\widetilde{F}_h^{X'}$ defined by (3.21) and $R_h^{X'}$ by (3.17). We split again

$$\left\|\widehat{F}_{h'}^{X'} - \widehat{F}_{h\vee h'}^{X'}\right\|_D^2 \mathbf{1}_{\Omega_{h,h'}} = \left\|\frac{\widetilde{F}_{h'}^{X'}}{R_{h'}^{X'}} - \frac{\widetilde{F}_{h\vee h'}^{X'}}{R_{h\vee h'}^{X'}}\right\|_D^2 \mathbf{1}_{\Omega_{h,h'}} \leq 4\left(T_{h'}^a + B_{h,h'} + \widetilde{T}_{h\vee h'} + T_{h\vee h'}^b\right),$$

where

$$\begin{aligned}
T_{h'}^a &= \left\|\frac{1}{R_{h'}^{X'}}\left(\widetilde{F}_{h'}^{X'} - \mathbb{E}_{X'}\left[\widetilde{F}_{h'}^{X'}\right]\right)\right\|_D^2 \mathbf{1}_{\Omega_{h,h'}}, \\
T_{h\vee h'}^b &= \left\|\frac{1}{R_{h\vee h'}^{X'}}\left(\widetilde{F}_{h\vee h'}^{X'} - \mathbb{E}_{X'}\left[\widetilde{F}_{h\vee h'}^{X'}\right]\right)\right\|_D^2 \mathbf{1}_{\Omega_{h,h'}}, \\
B_{h,h'} &= \frac{1}{\left(R_{h'}^{X'}\right)^2}\left\|\mathbb{E}_{X'}\left[\widetilde{F}_{h'}^{X'}\right] - \mathbb{E}_{X'}\left[\widetilde{F}_{h\vee h'}^{X'}\right]\right\|_D^2 \mathbf{1}_{\Omega_{h,h'}}, \\
\widetilde{T}_{h\vee h'} &= \left(\frac{1}{R_{h'}^{X'}} - \frac{1}{R_{h\vee h'}^{X'}}\right)^2 \left\|\mathbb{E}_{X'}\left[\widetilde{F}_{h\vee h'}^{X'}\right]\right\|_D^2 \mathbf{1}_{\Omega_{h,h'}}.
\end{aligned} \tag{3.28}$$

Thus, by subtracting $V(h')$ and taking the maximum over $h' \in \mathcal{H}_n$, we obtain

$$\begin{aligned}
A(h) &= \max_{h'\in\mathcal{H}_n}\left(\left\|\widehat{F}_{h'}^{X'} - \widehat{F}_{h\vee h'}^{X'}\right\|_D^2 - V(h')\right)_+ \\
&\leq \max_{h'\in\mathcal{H}_n}\left(4T_{h'}^a - \frac{V(h')}{3}\right)_+ + \max_{h'\in\mathcal{H}_n}\left(4T_{h\vee h'}^b - \frac{V(h')}{3}\right)_+ \\
&\quad + \max_{h'\in\mathcal{H}_n}\left(4\widetilde{T}_{h\vee h'} - \frac{V(h')}{3}\right)_+ \\
&\quad + 4\max_{h'\in\mathcal{H}_n} B_{h,h'} + \max_{h'\in\mathcal{H}_n}\left(\left\|\widehat{F}_{h'}^{X'} - \widehat{F}_{h\vee h'}^{X'}\right\|_D^2 \mathbf{1}_{\Omega_{h,h'}^c}\right).
\end{aligned} \tag{3.29}$$

It is not necessary to subtract $V(h')$ to the last two terms: we show below that the first one is of the order of the bias $h^{2\beta}$ and the second one is directly negligible. We now deal with each of the terms involving in (3.29) on the set $\{X' \in B\}$.

• **Upper bound for the term depending on $B_{h,h'}$.** We first use the definition of the set

$\Omega_{h,h'}$, and split the term to obtain the bias terms:

$$\begin{aligned}
\max_{h'\in\mathcal{H}_n} B_{h,h'} &\leq 4\max_{h'\in\mathcal{H}_n}\left\|\mathbb{E}_{X'}\left[\widetilde{F}_{h'}^{X'}\right]-\mathbb{E}_{X'}\left[\widetilde{F}_{h\vee h'}^{X'}\right]\right\|_D^2 \\
&= 4\max_{\substack{h'\in\mathcal{H}_n\\ h'\leq h}}\left\|\mathbb{E}_{X'}\left[\widetilde{F}_{h'}^{X'}\right]-\mathbb{E}_{X'}\left[\widetilde{F}_{h\vee h'}^{X'}\right]\right\|_D^2 \text{ since } B_{h,h'}=0 \text{ if } h'>h \\
&\leq 8\max_{\substack{h'\in\mathcal{H}_n\\ h'\leq h}}\left\{\left\|\mathbb{E}_{X'}\left[\widetilde{F}_{h'}^{X'}\right]-F^{X'}\right\|_D^2+\left\|F^{X'}-\mathbb{E}_{X'}\left[\widetilde{F}_{h\vee h'}^{X'}\right]\right\|_D^2\right\} \\
&\leq 8\left(\max_{\substack{h'\in\mathcal{H}_n\\ h'\leq h}} C_D^2(h')^{2\beta}+C_D^2h^{2\beta}\right)\leq 16C_D^2h^{2\beta}, \qquad (3.30)
\end{aligned}$$

thanks to Lemma 11.

• **Upper bound for the term depending on $\mathbf{1}_{\Omega^c_{h,h'}}$.** It is the second term which does not depend on $V(h')$:

$$\max_{h'\in\mathcal{H}_n}\left(\left\|\widehat{F}_{h'}^{X'}-\widehat{F}_{h\vee h'}^{X'}\right\|_D^2\mathbf{1}_{\Omega^c_{h,h'}}\right) \leq 2\max_{h'\in\mathcal{H}_n}\mathbf{1}_{\Omega^c_{h,h'}}\left(\left\|\widehat{F}_{h'}^{X'}\right\|_D^2+\left\|\widehat{F}_{h\vee h'}^{X'}\right\|_D^2\right).$$

Thus, since $|\widehat{F}_{h'}^{X'}(y)|\leq 1$ and $\left|\widehat{F}_{h'\vee h}^{X'}(y)\right|\leq 1$,

$$\mathbb{E}\left[\max_{h'\in\mathcal{H}_n}\left(\left\|\widehat{F}_{h'}^{X'}-\widehat{F}_{h\vee h'}^{X'}\right\|_D^2\mathbf{1}_{\Omega^c_{h,h'}}\mathbf{1}_B(X')\right)\right]\leq 4|D|\sum_{h'\in\mathcal{H}_n}\mathbb{P}\left(\Omega^c_{h,h'}\cap\{X'\in B\}\right).$$

Moreover,

$$\begin{aligned}
\mathbb{P}\left(\Omega^c_{h,h'}\cap\{X'\in B\}\right) &\leq \mathbb{P}\left(\left\{R_{h'}^{X'}<\frac{1}{2}\right\}\cap\{X'\in B\}\right)+\mathbb{P}\left(\left\{R_{h\vee h'}^{X'}<\frac{1}{2}\right\}\cap\{X'\in B\}\right) \\
&\leq \mathbb{P}\left(\left\{\left|R_{h'}^{X'}-1\right|>\frac{1}{2}\right\}\cap\{X'\in B\}\right)+\mathbb{P}\left(\left\{\left|R_{h'\vee h}^{X'}-1\right|>\frac{1}{2}\right\}\cap\{X'\in B\}\right).
\end{aligned}$$

Thus we apply Inequality (3.18) of Lemma 8, with $\eta=1/2$:

$$\sum_{h'\in\mathcal{H}_n}\mathbb{P}\left(\Omega^c_{h,h'}\cap\{X'\in B\}\right) \leq \sum_{h'\in\mathcal{H}_n}\left(2\exp\left(-\frac{n\varphi(h')}{8\left(\frac{M_2}{m_1^2}+\frac{C_K}{2m_1}\right)}\right)+2\exp\left(-\frac{n\varphi(h\vee h')}{8\left(\frac{M_2}{m_1^2}+\frac{C_K}{2m_1}\right)}\right)\right).$$

Recall now that thanks to H_{b_2}, $\varphi(h)\geq C_0\ln(n)/n$ for all $h\in\mathcal{H}_n$, with $C_0>16(M_2/m_1^2+C_K/2m_1)$. Use also H_{b_1} to deduce

$$\sum_{h'\in\mathcal{H}_n}\mathbb{P}\left(\Omega^c_{h,h'}\cap\{X'\in B\}\right) \leq 4\times n\times n^{-\frac{C_0}{8\left(\frac{M_2}{m_1^2}+\frac{C_K}{2m_1}\right)}}<\frac{4}{n}.$$

Thus, we have proved that

$$\mathbb{E}\left[\max_{h'\in\mathcal{H}_n}\left(\left\|\widehat{F}^{X'}_{h'}-\widehat{F}^{X'}_{h\vee h'}\right\|^2_D\mathbf{1}_{\Omega^c_{h,h'}}\right)\mathbf{1}_B(X')\right]\leq\frac{16|D|}{n}. \tag{3.31}$$

• **Upper bound for the term depending on $\widetilde{T}_{h,h'}$.** The definition of this term implies that

$$\begin{aligned}
\widetilde{T}_{h,h'} &= \left(\frac{R^{X'}_{h\vee h'}-R^{X'}_{h'}}{R^{X'}_{h'}R^{X'}_{h\vee h'}}\right)^2\left\|\mathbb{E}_{X'}\left[\widetilde{F}^{X'}_{h\vee h'}\right]\right\|^2_D\mathbf{1}_{\Omega_{h,h'}}\\
&\leq 16\left(R^{X'}_{h\vee h'}-R^{X'}_{h'}\right)^2\left\|\mathbb{E}_{X'}\left[\widetilde{F}^{X'}_{h\vee h'}\right]\right\|^2_D\\
&\leq 16|D|\left(R^{X'}_{h\vee h'}-R^{X'}_{h'}\right)^2\\
&\leq 32|D|\left\{\left(R^{X'}_{h\vee h'}-1\right)^2+\left(R^{X'}_{h'}-1\right)^2\right\},
\end{aligned}$$

using that $\mathbb{E}\left[\widetilde{F}^{X'}_{h\vee h'}\right]\leq 1$. We roughly bound the supremum over $h'\in\mathcal{H}_n$ by a sum over h' and use the last inequality:

$$\begin{aligned}
\mathbb{E}\left[\max_{h'\in\mathcal{H}_n}\left(4\widetilde{T}_{h\vee h'}-\frac{V(h')}{3}\right)_+\mathbf{1}_B(X')\right] &\leq 4\sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[\left(\widetilde{T}_{h\vee h'}-\frac{V(h')}{12}\right)_+\mathbf{1}_B(X')\right]\\
&\leq 4\sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[\left(32|D|\left\{\left(R^{X'}_{h\vee h'}-1\right)^2\right.\right.\right.\\
&\qquad\left.\left.\left.+\left(R^{X'}_{h'}-1\right)^2\right\}-\frac{V(h')}{12}\right)_+\mathbf{1}_B(X')\right]\\
&\leq 4\left\{\sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[\left(32|D|\left(R^{X'}_{h\vee h'}-1\right)^2-\frac{V(h')}{24}\right)_+\mathbf{1}_B(X')\right]\right.\\
&\qquad\left.+\sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[\left(32|D|\left(R^{X'}_{h'}-1\right)^2-\frac{V(h')}{24}\right)_+\mathbf{1}_B(X')\right]\right\}\\
&\leq 128|D|\left\{\sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[\left(\left(R^{X'}_{h\vee h'}-1\right)^2-\frac{V(h')}{768|D|}\right)_+\mathbf{1}_B(X')\right]\right.\\
&\qquad\left.+\sum_{h'\in\mathcal{H}_n}\mathbb{E}\left[\left(\left(R^{X'}_{h'}-1\right)^2-\frac{V(h')}{768|D|}\right)_+\mathbf{1}_B(X')\right]\right\}.
\end{aligned}$$

Then, Inequality (3.19) of Lemma 8 (with $\alpha=2$) proves that, on the set $\{X'\in B\}$, a.s.,

$$\mathbb{E}_{X'}\left[\left(\left(R^{X'}_{h'}-1\right)^2-V_R(h')\right)_+\right]\leq\min\left(\frac{4M_2}{m_1^2},\frac{64C_K^2}{m_1^2}\right)\frac{1}{n^2},$$

with $V_R(h')=\kappa_R\ln(n)/(n\varphi(h'))$ and $\kappa_R>\max(8M_2/m_1^2,128C_K^2/m_1^2C_0)$. Choosing $\kappa>$

$768|D|\kappa_R$ in the definition of $V(h')$ (see (3.24)) leads to $V(h')/768|D| \geq V_R(h')$, and hence we also have

$$\sum_{h'\in\mathcal{H}_n} \mathbb{E}\left[\left(\left(R_{h'}^{X'}-1\right)^2 - \frac{V(h')}{768|D|}\right)_+ \mathbf{1}_B(X')\right] \leq \min\left(\frac{4M_2}{m_1^2}, \frac{64C_K^2}{m_1^2}\right)\frac{1}{n},$$

thanks to Assumption H_{b_1}. Since $\varphi(h) \leq \varphi(h\vee h')$, $V(h') \geq V(h\vee h')$, the other term is bounded as follows

$$\mathbb{E}\left[\left(\left(R_{h\vee h'}^{X'}-1\right)^2 - \frac{V(h')}{768|D|}\right)_+ \mathbf{1}_B(X')\right] \leq \mathbb{E}\left[\left(\left(R_{h\vee h'}^{X'}-1\right)^2 - \frac{V(h\vee h')}{768|D|}\right)_+ \mathbf{1}_B(X')\right],$$

and same computations allow to deal with it. We thus deduce that

$$\mathbb{E}\left[\max_{h'\in\mathcal{H}_n}\left(4\widetilde{T}_{h\vee h'} - \frac{V(h')}{3}\right)_+ \mathbf{1}_B(X')\right] \quad\leq\quad 256|D|\min\left(\frac{4M_2}{m_1^2}, \frac{64C_K^2}{m_1^2}\right)\frac{1}{n}. \tag{3.32}$$

• **Upper bound for the terms depending on $T_{h'}^a$ or $T_{h'}^b$.** First, by definition of $\Omega_{h,h'}$, $T_{h'}^a \leq 4\|\widetilde{F}_{h'}^{X'} - \mathbb{E}_{X'}[\widetilde{F}_{h'}^{X'}]\|_D^2$. Furthermore, noticing that $\widetilde{F}_{h'}^{X'}$ belongs to $L^1(D)\cap L^2(D)$, we have:

$$\left\|\widetilde{F}_{h'}^{X'} - \mathbb{E}_{X'}[\widetilde{F}_{h'}^{X'}]\right\|_D^2 = \sup_{t\in\bar{S}_D(0,1)} \langle \widetilde{F}_{h'}^{X'} - \mathbb{E}_{X'}[\widetilde{F}_{h'}^{X'}], t\rangle_D^2, \tag{3.33}$$

where $\bar{S}_D(0,1)$ is a dense countable subset of the sphere $\{t \in L^1(D)\cap L^2(D), \|t\|_D = 1\}$ (such a set exists thanks to the separability of $L^2(D)$). Indeed, Cauchy-Schwarz inequality implies that, for all $t \in \bar{S}_D(0,1)$

$$\langle \widetilde{F}_{h'}^{X'} - \mathbb{E}_{X'}[\widetilde{F}_{h'}^{X'}], t\rangle_D^2 \leq \left\|\widetilde{F}_{h'}^{X'} - \mathbb{E}_{X'}[\widetilde{F}_{h'}^{X'}]\right\|_D^2.$$

Now let $(t_n)_{n\in\mathbb{N}^*}$ such that, for all $n \in \mathbb{N}^*$, $t_n \in \bar{S}_D(0,1)$ and

$$t_n \to_{n\to+\infty} (\widetilde{F}_{h'}^{X'} - \mathbb{E}_{X'}[\widetilde{F}_{h'}^{X'}])/\left\|\widetilde{F}_{h'}^{X'} - \mathbb{E}_{X'}[\widetilde{F}_{h'}^{X'}]\right\|_D^2,$$

we have $\langle \widetilde{F}_{h'}^{X'} - \mathbb{E}_{X'}[\widetilde{F}_{h'}^{X'}], t_n\rangle_D^2 \to_{n\to+\infty} \left\|\widetilde{F}_{h'}^{X'} - \mathbb{E}_{X'}[\widetilde{F}_{h'}^{X'}]\right\|_D^2$ which concludes the proof of (3.33). Moreover, we write the scalar product $\langle \widetilde{F}_{h'}^{X'} - \mathbb{E}_{X'}[\widetilde{F}_{h'}^{X'}], t\rangle_D \mathbf{1}_B(X') = \nu_{n,h}(t)$, for $t \in \bar{S}_D(0,1)$, where

$$\begin{aligned}\nu_{n,h}(t) \quad &= \quad \frac{1}{n}\sum_{i=1}^n \psi_{t,h}(X_i,Y_i) - \mathbb{E}_{X'}\left[\psi_{t,h}(X_i,Y_i)\right] \\ &\qquad \text{with } \psi_{t,h}(X_i,Y_i) = \frac{K_h(d(X_i,X'))}{\mathbb{E}_{X'}[K_h(d(X_i,X'))]}\langle \mathbf{1}_{[Y_i;\infty[}, t\rangle_D \mathbf{1}_B(X').\end{aligned} \tag{3.34}$$

Consequently,

$$\mathbb{E}\left[\max_{h'\in\mathcal{H}_n}\left(4T_{h'}^a - \frac{V(h')}{3}\right)_+ \mathbf{1}_B(X')\right] \leq 16\sum_{h'\in\mathcal{H}_n} \mathbb{E}\left[\left(\sup_{t\in\bar{S}_D(0,1)} \nu_{n,h'}^2(t) - \frac{V(h')}{48}\right)_+ \mathbf{1}_B(X')\right].$$

We use the following lemma, which permits to control the empirical process defined by (3.34).

Lemma 13. *Under the assumptions of Theorem 4, for* $\delta_0 > \max(3528 C_K^2 |D|/M_2 C_0, 12)$, *there exists a constant* $C > 0$ *(depending only on* m_1, M_2, δ_0, C_0 *and* $|D|$*) such that*

$$\sum_{h \in \mathcal{H}_n} \mathbb{E}\left[\left(\sup_{t \in \bar{S}_D(0,1)} \nu_{n,h}^2(t) - 6\delta_0 \frac{|D| M_2}{m_1^2} \frac{\ln(n)}{n\varphi(h)} \right)_+ \mathbf{1}_B(X') \right] \leq \frac{C}{n}.$$

Choosing $\kappa > 288\delta_0 |D| M_2/m_1^2$ in the definition of $V(h')$ (see (3.24)) leads to $V(h')/48 \geq 6\delta_0 |(D|M_2/m_1^2) \ln(n)/(n\varphi(h))$. This proves that

$$\mathbb{E}\left[\max_{h' \in \mathcal{H}_n} \left(4T_{h'}^a - \frac{V(h')}{3} \right)_+ \mathbf{1}_B(X') \right] \leq \frac{16C}{n}. \tag{3.35}$$

Recall finally that $V(h') \geq V(h \vee h')$, similar computations allow to also obtain the same bound for $\mathbb{E}[\max_{h' \in \mathcal{H}_n} \left(4T_{h \vee h'}^b - V(h')/3\right)_+]$.

Gathering Inequalities (3.30), (3.31), (3.32), and (3.35) in Inequality (3.29) completes the proof of Lemma 12.

□

Proof of Lemma 13 (concentration of the empirical process)

The aim is to control the deviations of the supremum of the empirical process $\nu_{n,h}$ defined by (3.34). Since it is centred and bounded, the guiding idea is to apply Talagrand's Inequality (Lemma 24 p. 180).

We first compute H^ν, M^ν and v^ν, involved in Lemma 24.

- For M^ν, let $t \in \bar{S}_D(0,1)$, $x \in \mathbb{H}$ and $y \in \mathbb{R}$ be fixed. By the Cauchy-Schwarz Inequality,

$$|\psi_{t,h}(x,y)| \leq |D| \|t\|_D \frac{\|K_h\|_{L^\infty(\mathbb{R})}}{m_1 h^{-1} \varphi(h)} \leq \frac{|D| C_K}{m_1 \varphi(h)} := M_1^\nu,$$

  thanks to (3.16).
- For H^ν, recall that

$$\mathbb{E}_{X'}\left[\sup_{t \in \bar{S}_D(0,1)} \nu_{n,h}^2(t) \right] = \mathbb{E}_{X'}\left[\left\| \widetilde{F}_{h'}^{X'} - \mathbb{E}_{X'}[\widetilde{F}_{h'}^{X'}] \right\|_D^2 \right] \leq |D| \frac{M_2}{m_1^2} \frac{1}{n\varphi(h)} := (H^\nu)^2$$

  a.s. on the set $\{X' \in B\}$ with the same computation as for the variance term, see Lemma 9.

– For v^ν, we also fix $t \in \bar{S}_D(0,1)$, and compute,

$$\begin{aligned}\mathrm{Var}_{X'}\left(\psi_{t,h}(X_1,Y_1)\right) &\leq \mathbb{E}_{X'}\left[\psi^2_{t,h}(X_1,Y_1)\right], \\ &= \mathbb{E}_{X'}\left[\left(\int_D \mathbf{1}_{Y_1\leq y}t(y)dy\right)^2 \frac{K_h^2(d(X_1,X'))}{\left(\mathbb{E}_{X'}\left[K_h(d(X_1,X'))\right]\right)^2}\right]\mathbf{1}_B(X').\end{aligned}$$

The integral is controlled with the Cauchy-Schwarz Inequality: $(\int_D \mathbf{1}_{Y_1\leq y}t(y)dy)^2 \leq |D|\|t\|_D^2 = |D|$, and the other quantity has already been bounded: we obtain

$$\mathrm{Var}_{X'}\left(\psi_{t,h}(X_1,Y_1)\right) \leq v^\nu := \frac{|D|M_2}{m_1^2\varphi(h)}.$$

Then, Lemma 24 gives, for $\delta > 0$,

$$\begin{aligned}\mathbb{E}\left[\left(\sup_{t\in\bar{S}_D(0,1)} \nu^2_{n,h}(t) - 2(1+2\delta)\left(H^\nu\right)^2\right)_+ \mathbf{1}_B(X')\right] \leq C\left\{\frac{|D|M_2}{m_1^2}\frac{1}{n\varphi(h)}\exp\left(-\frac{\delta}{6}\right)\right. \\ \left. + \frac{1}{C^2(\delta)}\frac{C_K^2|D|^2}{m_1^2}\frac{1}{n^2\varphi^2(h)}\exp\left(-\frac{1}{21\sqrt{2}}C(\delta)\sqrt{\delta}\frac{\sqrt{M_2}\sqrt{n\varphi(h)}}{\sqrt{|D|}C_K}\right)\right\}.\end{aligned}$$

We choose $\delta = \delta_0\ln(n)$, for a δ_0 large enough, and given below. We compute the order of magnitude of the last upper bound, using Assumptions H_{b_1} and H_{b_2}. Recall that they imply $\sum_{h\in\mathcal{H}_n} 1/\varphi(h) \leq n^2/C_0\ln(n)$ and $\sum_{h\in\mathcal{H}_n} 1/\varphi^2(h) \leq n^3/C_0^2\ln^2(n)$. First,

$$\sum_{h\in\mathcal{H}_n}\frac{1}{n\varphi(h)}\exp\left(-\frac{\delta}{6}\right) = \frac{1}{n^{1+\delta_0/6}}\sum_{h\in\mathcal{H}_n}\frac{1}{\varphi(h)} \leq \frac{1}{C_0 n^{\delta_0/6-1}\ln(n)} \leq \frac{1}{C_0 n},$$

as soon as $\delta_0 \geq 12$ since we can reasonably assume $n \geq 3$. Then, $C(\delta_0\ln(n)) = \sqrt{1+\delta_0\ln(n)} - 1 \geq 1$, if $\delta_0\ln(n) \geq 3$, that is $\ln(n) \geq 3/\delta_0$. This is satisfied since $\delta_0 > 12$ and $n \geq 2$. Hence

$$\begin{aligned}&\frac{1}{n^2C^2(\delta)}\sum_{h\in\mathcal{H}_n}\frac{1}{\varphi^2(h)}\exp\left(-\frac{1}{21\sqrt{2}}C(\delta_0\ln(n))\sqrt{\delta_0\ln(n)}\frac{\sqrt{M_2}}{\sqrt{|D|}C_K}\sqrt{n\varphi(h)}\right) \\ &\quad\leq \frac{1}{n^2}\sum_{h\in\mathcal{H}_n}\frac{1}{\varphi^2(h)}\exp\left(-\frac{1}{21\sqrt{2}}\sqrt{\delta_0\ln(n)}\frac{\sqrt{M_2}}{\sqrt{|D|}C_K}\sqrt{n\varphi(h)}\right) \\ &\quad\leq \frac{1}{n^2}\sum_{h\in\mathcal{H}_n}\frac{1}{\varphi^2(h)}\exp\left(-\frac{\sqrt{C_0}}{21\sqrt{2}}\sqrt{\delta_0}\frac{\sqrt{M_2}}{\sqrt{|D|}C_K}\ln(n)\right) \\ &\quad= n^{-2-\frac{\sqrt{C_0M_2\delta_0}}{21\sqrt{2|D|}C_K}}\sum_{h\in\mathcal{H}_n}\frac{1}{\varphi^2(h)} \leq \frac{1}{C_0^2\ln^2(n)}n^{-\frac{\sqrt{C_0M_2\delta_0}}{21\sqrt{2|D|}C_K}+1} \leq \frac{1}{C_0^2 n},\end{aligned}$$

as soon as $\sqrt{C_0M_2\delta_0}/(21\sqrt{2|D|}C_K) - 1 > 1$ that is $\delta_0 > 3528C_K^2|D|/C_0M_2$ with $C > 0$

depending only on m_1, M_2, δ_0, C_0 and $|D|$. This shows that

$$\sum_{h\in\mathcal{H}_n}\mathbb{E}\left[\left(\sup_{t\in\bar{S}_D(0,1)}\nu_{n,h}^2(t)-2(1+2\delta_0\ln(n))\left(H^\nu\right)^2\right)_+\mathbf{1}_B(X')\right]\leq\frac{C}{n}, \tag{3.36}$$

for $(H^\nu)^2=|D|(M_2/m_1^2)/(n\varphi(h))$ and $C>0$ depending only on m_1, M_2, δ_0, C_0 and $|D|$. Since

$$2(1+2\delta_0\ln(n))\left(H^\nu\right)^2\leq 6\delta_0\frac{|D|M_2}{m_1^2}\frac{\ln(n)}{n\varphi(h)},$$

Inequality (3.36) is also satisfied when we replace $2(1+2\delta_0\ln(n))\left(H^\nu\right)^2$ by this upper bound. Thus, the proof of Lemma 13 is completed.

□

3.7.4 Proof of Theorem 4

Let Λ be the set

$$\Lambda=\bigcap_{h\in\mathcal{H}_n}\left\{\left|\frac{\widehat{\varphi}(h)}{\varphi(h)}-1\right|<\frac{1}{2}\right\}.$$

We split the loss function of the estimator

$$\left\|\widehat{F}^{X'}_{\widehat{h}}-F^{X'}\right\|_D^2\leq\left\|\widehat{F}^{X'}_{\widehat{h}}-F^{X'}\right\|_D^2\left(\mathbf{1}_\Lambda+\mathbf{1}_{\Lambda^c}\right).$$

We will argue as follows: first, on the set Λ, $\widehat{\varphi}(h)$ is close to $\varphi(h)$, and we use the same arguments as for (3.25). Second, the probability of the set Λ^c is negligible. Let us prove these two claims.

• **Upper bound for $\|\widehat{F}^{X'}_{\widehat{h}}-F^{X'}\|_D^2\mathbf{1}_\Lambda$.** It follows from the same arguments as in the beginning of the proof of Inequality (3.25) (see Section 3.7.3), that

$$\left\|\widehat{F}^{X'}_{\widehat{h}}-F^{X'}\right\|_D^2\mathbf{1}_\Lambda\leq\left\{6\widehat{A}(h)+6\widehat{V}(h)+3\left\|\widehat{F}^{X'}_{h}-F^{X'}\right\|_D^2\right\}\mathbf{1}_\Lambda.$$

Note that

$$\begin{aligned}\widehat{A}(h) &= \max_{h'\in\mathcal{H}_n,\,\widehat{V}(h')<\infty}\left\{\left\|\widehat{F}^{X'}_{h}-\widehat{F}^{X'}_{h\vee h'}\right\|_D^2-V(h')+\left(V(h')-\widehat{V}(h')\right)\right\}_+\\ &\leq \max_{h'\in\mathcal{H}_n,\,\widehat{V}(h')<\infty}\left\{\left\|\widehat{F}^{X'}_{h}-\widehat{F}^{X'}_{h\vee h'}\right\|_D^2-V(h')\right\}+\max_{h'\in\mathcal{H}_n,\,\widehat{V}(h')<\infty}\left(V(h')-\widehat{V}(h')\right)_+\\ &\leq A(h)+\max_{h'\in\mathcal{H}_n}\left(V(h')-\widehat{V}(h')\right)_+.\end{aligned}$$

We obtain the following decomposition:

$$\left\|\widehat{F}^{X'}_{\widehat{h}} - F^{X'}\right\|^2_D \mathbf{1}_\Lambda \leq \left\{6A(h) + 6V(h) + 3\left\|\widehat{F}^{X'}_{h} - F^{X'}\right\|^2_D + 6\max_{h'\in\mathcal{H}_n}\left(V(h') - \widehat{V}(h')\right)_+ + 6\left(\widehat{V}(h) - V(h)\right)\right\}\mathbf{1}_\Lambda. \tag{3.37}$$

For $h' \in \mathcal{H}_n$, such that $\widehat{V}(h') < \infty$, we have

$$V(h') - \widehat{V}(h') = \kappa\frac{\ln(n)}{n}\left(\frac{1}{\varphi(h')} - \frac{3}{2}\frac{1}{\widehat{\varphi}(h')}\right).$$

But on the set Λ, for any $h' \in \mathcal{H}_n$, $|\widehat{\varphi}(h') - \varphi(h')| < \varphi(h')/2$. In particular, we thus have $\widehat{\varphi}(h') - \varphi(h') < \varphi(h')/2$, that is $\widehat{\varphi}(h') < (3/2)\varphi(h')$. This proves that $V(h') - \widehat{V}(h') < 0$, and hence

$$\max_{h'\in\mathcal{H}_n}\left(V(h') - \widehat{V}(h')\right)_+ = 0.$$

Moreover, on Λ, we also have, for $h \in \mathcal{H}_n$, $\varphi(h) - \widehat{\varphi}(h) < \varphi(h)/2$, that is $2/\varphi(h) > 1/\widehat{\varphi}(h)$. Thus,

$$\widehat{V}(h) - V(h) = \kappa\frac{\ln(n)}{n}\left(\frac{3}{2}\frac{1}{\widehat{\varphi}(h)} - \frac{1}{\varphi(h)}\right) \leq \kappa\frac{\ln(n)}{n}2\frac{1}{\varphi(h)} = 2V(h).$$

Gathering the two bounds in (3.37) leads to

$$\left\|\widehat{F}^{X'}_{\widehat{h}} - F^{X'}\right\|^2_D \mathbf{1}_\Lambda \leq 6A(h) + 8V(h) + 3\left\|\widehat{F}^{X'}_{h} - F^{X'}\right\|^2_D.$$

We thus obtain, like in the proof of Inequality (3.25),

$$\mathbb{E}\left[\left\|\widehat{F}^{X'}_{\widehat{h}} - F^{X'}\right\|^2_D \mathbf{1}_\Lambda \mathbf{1}_B(X')\right] \leq 8V(h) + 3C\left(h^{2\beta} + \frac{1}{n\varphi(h)}\right).$$

• **Upper bound for $\|\widehat{F}^{X'}_{\widehat{h}} - F^{X'}\|^2_D \mathbf{1}_{\Lambda^c}$.** We roughly bound

$$\left\|\widehat{F}^{X'}_{\widehat{h}} - F^{X'}\right\|^2_D \mathbf{1}_{\Lambda^c} \leq 2\left(\left\|\widehat{F}^{X'}_{\widehat{h}}\right\|^2_D + \left\|F^{X'}\right\|^2_D\right)\mathbf{1}_{\Lambda^c} \leq 4|D|\mathbf{1}_{\Lambda^c}.$$

It remains to control $\mathbb{P}(\Lambda^c)$:

$$\begin{aligned}\mathbb{P}(\Lambda^c) &\leq \sum_{h\in\mathcal{H}_n}\mathbb{P}\left(|\widehat{\varphi}(h) - \varphi(h)| \geq \frac{\varphi(h)}{2}\right), \\ &= \sum_{h\in\mathcal{H}_n}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{d(0,X_i)\leq h\}} - \mathbb{E}\left[\mathbf{1}_{\{d(0,X_i)\leq h\}}\right]\right| \geq \frac{\varphi(h)}{2}\right).\end{aligned}$$

We apply Bernstein's Inequality (Lemma 22), with $T_i = \mathbf{1}_{\{d(0,X_i)\leq h\}}$ and $\eta = \varphi(h)/2$. Since

$0 \leq T_i \leq 1$, we set $b_0 = 1$, and $v^2 = \mathrm{Var}(T_1) = \varphi(h)(1-\varphi(h))$. We derive

$$\begin{aligned}\mathbb{P}(\Lambda^c) &\leq 2\sum_{h\in\mathcal{H}_n} \exp\left(-\frac{n\varphi^2(h)/8}{\varphi(h)(1-\varphi(h))+\varphi(h)/2}\right)\\ &= 2\sum_{h\in\mathcal{H}_n} \exp\left(-\frac{n\varphi(h)}{8(1-\varphi(h))+4}\right) \leq 2\sum_{h\in\mathcal{H}_n} \exp\left(-\frac{n\varphi(h)}{12}\right).\end{aligned}$$

We thus obtain

$$\mathbb{E}\left[\left\|\widehat{F}^{X'}_{\widetilde{h}} - F^{X'}\right\|^2_D \mathbf{1}_{\Lambda^c}\mathbf{1}_B(X')\right] \leq 8|D|\sum_{h\in\mathcal{H}_n} \exp\left(-\frac{n\varphi(h)}{12}\right) \leq 8|D|n^{1-C_0/12}.$$

thanks to Assumptions H_{b1} and H_{b2}. Taking $C_0 > 24$, we have $n^{1-C_0/12} < n^{-1}$ which ends the proof.

□

3.7.5 Proof of Theorem 5

Proof of (a)

We have to compute the convergence under three regularity assumptions ($H_{X,L}$, $H_{X,M}$ and $H_{X,F}$), and for the two criteria (pointwise and integrated). It follows from (3.6) and (3.7) of Theorem 3 that the risk of the estimator is bounded, up to a multiplicative constant, by

$$\widetilde{R}(h) = h^{2\beta} + 1/(n\varphi^{x_0}(h)),$$

with $x_0 = 0$ for the integrated risk. To obtain the convergence rates, it is thus sufficient to compute the bandwidth h which minimizes the bound $\widetilde{R}(h)$ when assuming H_X, or at least to choose it well.

Convergence rate under Assumption $H_{X,L}$ With the lower bound on φ^{x_0} of $H_{X,L}$, the quantity $\widetilde{R}(h)$ and thus also the risks are upper bounded by a quantity with order of magnitude

$$R(h) := h^{2\beta} + h^{-\gamma}\exp\left(c_2 h^{-\alpha}\right) n^{-1}.$$

Choosing the bandwidth h_0 such that

$$h_0 = \left(\frac{\ln n}{c_2} - \kappa \ln\left(\frac{\ln n}{c_2}\right)\right)^{-1/\alpha},$$

with $\kappa := c_2^{-1}(\gamma/\alpha + 2\beta/\alpha)$ ends the proof, since $R(h_0)$ has the announced order.

□

Convergence rate under Assumption $H_{X,M}$ ***or*** $H_{X,F}$ First assume $H_{X,M}$. The risk is bounded (up to a multiplicative constant) by the quantity

$$R(h) = h^{2\beta} + n^{-1}h^{-\gamma_1}\exp\left(c_2 \ln^{\alpha}(1/h)\right).$$

Choosing

$$h_0 = \exp\left(- \left(\frac{1}{c_2} \ln n - c_2^{-(\alpha+1)/\alpha} (2\beta - \gamma_1) \ln^{1/\alpha} n \right)_+^{1/\alpha} \right),$$

leads to what we want to prove, that is

$$R(h_0) \leq C \exp\left(- \frac{2\beta}{c_2^{1/\alpha}} \ln^{1/\alpha} n \right). \tag{3.38}$$

Indeed, if $n > \exp\left(c_2^{1/(1-\alpha)} (2\beta - \gamma_1)_+^{\alpha/(\alpha-1)} \right)$, we have

$$\frac{1}{c_2} \ln n - c_2^{-(\alpha+1)/\alpha} (2\beta - \gamma_1) \ln^{1/\alpha} n > 0$$

and

$$h_0 = \exp\left(- \left(\frac{1}{c_2} \ln n - c_2^{-(\alpha+1)/\alpha} (2\beta - \gamma_1) \ln^{1/\alpha} n \right)^{1/\alpha} \right).$$

Now, let $s_n := \exp\left(- \frac{2\beta}{c_2^{1/\alpha}} \ln^{1/\alpha} n \right)$, we have

$$\frac{h_0^{2\beta}}{s_n} = \exp\left(\frac{2\beta}{c_2^{1/\alpha}} \ln^{1/\alpha} n \left(1 - \left(1 - c_2^{-1/\alpha} (2\beta - \gamma_1) \ln^{1/\alpha - 1} \right)^{1/\alpha} \right) \right) \to_{n\to\infty} 1,$$

using that,

$$(1 - c_2^{-1/\alpha} (2\beta - \gamma_1) \ln^{1/\alpha - 1} n)^{1/\alpha} = 1 - \frac{1}{\alpha} c_2^{-1/\alpha} (2\beta - \gamma_1) \ln^{1/\alpha - 1} n + o(\ln^{1/\alpha - 1} n).$$

Then $h_0^{2\beta} \leq C s_n$. We deal similarly with the second term of $R(h)$,

$$\begin{aligned} &\frac{n^{-1} h_0^{-\gamma_1} \exp(c_2 \ln^\alpha (1/h))}{s_n} \\ &\quad = \exp\left(\gamma_1 c_2^{-1/\alpha} \ln^{1/\alpha} n \left(1 - \left(1 - c_2^{-1/\alpha} (2\beta - \gamma_1) \ln^{1/\alpha - 1} \right)^{1/\alpha} \right) \right) \\ &\qquad \to_{n\to\infty} 1, \end{aligned}$$

which leads to (3.38), and ends the computation of the rate under $H_{X,M}$.

When assuming $H_{X,F}$, the optimal h can be computed: the one which minimizes $R(h) = h^{2\beta} + h^{-\gamma}$ has the order $n^{1/(2\beta+\gamma)}$ and immediatly gives $R(h) \leq C n^{-2\beta/(2\beta+\gamma)}$.

□

Proof of (b)

The proof comes down to the proof of (a) since Theorem 3 gives a bound of the risks of $\widehat{F}_{\widehat{h}}$ with the form $\min_h \widetilde{R}(h)$ ($\widetilde{R}(h)$ defined in the proof of (a)). The computation of the (a)

bound for this minimum has thus been done in the previous section.

□

3.7.6 Proof of Theorem 6

Proof of (i), under Assumption $H_{X,L}$

The proof is based on the general reduction scheme described in Section 2.2 of Tsybakov (2009). Let $x_0 \in \mathbb{H}$ be fixed and $r_n = (\ln(n))^{-\beta/\alpha}$ the rate of convergence. We define two functions F_0 and F_1, called hypotheses, such that

(A) F_l belongs to $\mathcal{F}_\beta$, for $l = 0, 1$,

(B) $\|F_0^{x_0} - F_1^{x_0}\|_D^2 \geq cr_n$ for a constant $c > 0$,

(C) $K(\mathbb{P}_1^{\otimes n}, \mathbb{P}_0^{\otimes n}) \leq \alpha$ for a real number $\alpha < \infty$ (which does not depend on x_0), where $\mathbb{P}_0^{\otimes n}$ (resp. $\mathbb{P}_1^{\otimes n}$) is the probability distribution of a sample $(X_{0,i}, Y_{0,i})_{i=1,\dots,n}$ (resp. $(X_{1,i}, Y_{1,i})_{i=1,\dots,n}$) for which the conditional c.d.f. of $Y_{0,i} \in \mathbb{R}$ given $X_{0,i} \in \mathbb{H}$ (resp. of $Y_{1,i}$ given $X_{1,i}$) is F_0 (resp. F_1). $K(P, Q)$ is the Kullback-Leibler divergence between two probability distributions P and Q: $K(P, Q) = \int \ln(dP/dQ)dP$ if $P << Q$, and $K(P, Q) = +\infty$ otherwise.

Once these functions F_0 and F_1, verifying (A), (B) and (C), are constructed (which is done in the next paragraph), we can see that the proof is complete. Indeed, remark that, for all estimator $\widehat{F}^{x_0}$ constructed with the sample $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, for all $F \in \mathcal{F}_\beta$, we have, by the Markov Inequality,

$$\mathbb{E}_F\left[\|\widehat{F}^{x_0} - F^{x_0}\|_D^2\right] \geq \frac{c}{2} r_n \mathbb{P}_F\left(\|\widehat{F}^{x_0} - F^{x_0}\|_D^2 \geq \frac{c}{2} r_n\right),$$

where we recall that $\mathbb{E}_F$ (resp. $\mathbb{P}_F$) is the expectation (resp. probability) with respect to the law of $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ when, for all $i = 1, \dots, n$, the conditional expectation of Y_i given $X_i = x_0$ is F^{x_0}. Moreover, Hypothesis (A) implies that

$$\sup_{F \in \mathcal{F}_\beta} \mathbb{P}_F\left(\|\widehat{F}^{x_0} - F^{x_0}\|_D^2 \geq \frac{c}{2} r_n\right) \geq \max_{j=0,1} \mathbb{P}_F\left(\|\widehat{F}^{x_0} - F_j^{x_0}\|_D^2 \geq \frac{c}{2} r_n\right).$$

Now, let ψ^* be the minimum distance test defined by

$$\psi^* := \arg\min_{j=0,1} \|\widehat{F}^{x_0} - F_j^{x_0}\|_D^2,$$

if, for $j = 0, 1$, $\psi^* \neq j$, then

$$\|\widehat{F}^{x_0} - F_j^{x_0}\|_D^2 \geq \|\widehat{F}^{x_0} - F_{1-j}^{x_0}\|_D^2. \tag{3.39}$$

Then by the triangle inequality

$$\begin{aligned} \|\widehat{F}^{x_0} - F_j^{x_0}\|_D^2 &\geq \|F_{1-j}^{x_0} - F_j^{x_0}\|_D^2 - \|\widehat{F}^{x_0} - F_{1-j}^{x_0}\|_D^2 \\ &\geq cr_n - \|\widehat{F}^{x_0} - F_j^{x_0}\|_D^2, \end{aligned}$$

where the last inequality comes from Hypothesis (B) and Equation (3.39). Consequently

$$\psi^* \neq j \Rightarrow \|\widehat{F}^{x_0} - F_j^{x_0}\|_D^2 \geq \frac{c}{2} r_n.$$

Then,

$$\mathbb{P}_F\left(\|\widehat{F}^{x_0} - F_j^{x_0}\|_D^2 \geq \frac{c}{2} r_n\right) \geq \mathbb{P}_F(\psi^* \neq j)$$

and

$$\inf_{\widehat{F}} \max_{j=0,1} \mathbb{P}_F\left(\|\widehat{F}^{x_0} - F_j^{x_0}\|_D^2 \geq \frac{c}{2} r_n\right) \geq \inf_{\psi} \max_{j=0,1} \mathbb{P}_F(\psi \neq j),$$

where the second infimum is taken over all tests ψ calculated from the sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ (we recall that a test is a measurable application from $(\mathbb{H} \times R)^n$ to $\{0, 1\}$). Then, the following theorem, together with Hypothesis (C) allows us to conclude the proof.

Theorem 7. (Tsybakov, 2009, Theorem 2.2 (iii), p.90) *Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two probability measures on a measurable space $(E, \mathcal{B})$. If there exists a real number $\alpha < \infty$ such that*

$$K(\mathbb{P}_0^{\otimes n}, \mathbb{P}_1^{\otimes n}) \leq \alpha,$$

then

$$\inf_{\psi} \max_{j=0,1} \mathbb{P}_F(\psi \neq j) \geq \max\{e^{-\alpha}/4; (1 - \sqrt{\alpha/2})/2\},$$

where the infimum is taken over all tests.

In the sequel, we define F_0 and F_1 and check the three conditions (A), (B) and (C).

Construction of F_0 and F_1 and of the associated samples For $(x, y) \in \mathbb{H} \times \mathbb{R}$, let F_0^x be the c.d.f. of the uniform distribution on D, that is $F_0^x(y) = \frac{y}{|D|}\mathbf{1}_{y \in D} + \mathbf{1}_{y > \sup D}$. Choose a real random variable Y_0 with a uniform distribution $\mathbb{P}_{\mathcal{U}_D}$ on the compact set D, and take any process X_0 on $\mathbb{H}$, independent on Y_0, with distribution $\mathbb{P}_X$ verifying $H_{X,L}$. For the second function, set

$$F_1^x(y) = F_0^x(y) + L\eta_n^\beta H\left(\frac{\|x - x_0\|}{\eta_n}\right) \int_{-\infty}^{y} \psi(t)dt,$$

where

- $\psi : \mathbb{R} \to \mathbb{R}$ is a non-zero continuous function with support D with $\int_{\mathbb{R}} \psi(t)dt = 0$,
- $H : \mathbb{R}_+ \to \mathbb{R}_+$ is a function supported on $[0; 1]$ such that $|H(u) - H(v)| \leq |u - v|^\beta$, for any $(u, v) \in \mathbb{R}_+^2$,
- L is a real number such that $0 < L < 1/(\sup_{n \in \mathbb{N}^*}\{\eta_n^\beta\}|D|\|K\|L^\infty(\mathbb{R})\|\psi\|_{L^\infty(\mathbb{R})})$,
- η_n is a non-negative real number such that

$$\eta_n^{2\beta} \geq c_{(B)} r_n \text{ and } \eta_n^{2\beta} \varphi^{x_0}(\eta_n) \leq \frac{c_{(C)}}{n}, \tag{3.40}$$

for two constants $c_{(B)} > 0$ and $c_{(C)} > 0$.

From $H_{X,L}$, a positive number η_n for which the properties above hold is given by

$$\eta_n = \left(\frac{\ln n - ((2\beta+\gamma)/\alpha) \ln \ln n}{C_1} \right)^{-1/\alpha}. \tag{3.41}$$

We also choose a variable Y_1, such that, for any $x \in \mathbb{H}$, the conditional distribution of Y_1 given $X_0 = x$ is characterized by the c.d.f. F_1^x. The notation $\mathbb{P}_1$ is the distribution of (X_0, Y_1).

Checks of the conditions (A) to (C)

Check (A): belonging to the space $\mathcal{F}_\beta$ For any $x \in \mathbb{H}$, the function F_0^x is a c.d.f. by construction (it does not depend on x and is simply the c.d.f. of the uniform distribution on D), and $\|F_0^x - F_0^{x'}\|^2 = 0$ $(x, x' \in \mathbb{H})$. Thus, F_0 belongs to $\mathcal{F}_\beta$.

Let $x \in \mathbb{H}$ be fixed. The function $y \mapsto F_1^x(y)$ is continuous, with limit 0 when y goes to $-\infty$ (recall that D is a bounded set), and 1 when y goes to $+\infty$ (since $\int_{\mathbb{R}} \psi(t)dt = 0$). If $y \notin \bar{D}$, $(F_1^x)'(y) = 0$ (the support of ψ is included in D) and if $y \in \mathring{D}$,

$$(F_1^x)'(y) = \frac{1}{|D|} + L\eta_n^\beta H\left(\frac{\|x - x_0\|}{\eta_n} \right) \psi(y) \geq \frac{1}{|D|} - L\eta_n^\beta \|H\|_{L^\infty(\mathbb{R})} \|\psi\|_{L^\infty(\mathbb{R})} > 0,$$

thanks to the definition of L above. Thus F_1^x is increasing, and F_1^x is a conditional distribution function. Moreover, for any $x, x' \in \mathbb{H}$, denoting by $I_\psi = \int_D (\int_{-\infty}^y \psi(t)dt)^2 dy$,

$$\begin{aligned} \|F_1^x - F_1^{x'}\|_D^2 &= L^2 \eta_n^{2\beta} I_\psi \left(H\left(\frac{\|x - x_0\|}{\eta_n} \right) - H\left(\frac{\|x' - x_0\|}{\eta_n} \right) \right)^2 \\ &\leq L^2 \eta_n^{2\beta} I_\psi \left(\frac{\|x - x_0\|}{\eta_n} - \frac{\|x' - x_0\|}{\eta_n} \right)^{2\beta} \leq L^2 I_\psi \|x - x'\|^{2\beta}, \end{aligned}$$

thanks to the regularity property of the function H. Therefore, F_1 also belongs to $\mathcal{F}_\beta$.

Check (B): condition on the loss $\|F_0^{x_0} - F_1^{x_0}\|_D^2$ We have, thanks to the lower bound for η_n,

$$\|F_1^{x_0} - F_0^{x_0}\|_D^2 = L^2 \eta_n^{2\beta} H^2(0) I_\psi \geq L^2 H^2(0) I_\psi c_{(C)} r_n.$$

Check (C): Upper bound for the Kullback divergence $K(P_1^{\otimes n}, P_0^{\otimes n})$ In a first step, we prove that the measure $\mathbb{P}_1$ is absolutely continuous with respect to $\mathbb{P}_0$, and compute the Radon-Nikodym derivative. First, notice that

$$F_1^x(y) = \int_{-\infty}^y \frac{1}{|D|} \mathbf{1}_D(t) + L\eta_n^\beta H\left(\frac{\|x - x_0\|}{\eta_n} \right) \psi(t)dt.$$

Therefore, keeping in mind that $\int_{\mathbb{R}} \psi(t)dt = \int_D \psi(t)dt = 0$, the conditional distribution of Y_1 given $X_0 = x$ admits a density with respect to the Lebesgue measure on D given by

$$f_1^x(y) = \left(\frac{1}{|D|} + L\eta_n^\beta H\left(\frac{\|x - x_0\|}{\eta_n} \right) \psi(y) \right) \mathbf{1}_D(y).$$

We can thus compute the distribution $\mathbb{P}_1$ of the random couple (X_0, Y_1). For any test function Φ on $\mathbb{H} \times \mathbb{R}$,

$$\begin{aligned}
\int_{\mathbb{H}\times\mathbb{R}} \Phi(x,y) d\mathbb{P}_1(x,y) &= \mathbb{E}\left[\Phi(X_0, Y_1)\right] = \mathbb{E}\left[\mathbb{E}\left[\Phi(X_0, Y_1)|X_0\right]\right] \\
&= \int_{\mathbb{H}} \mathbb{E}\left[\Phi(x, Y_1)|X_0 = x\right] d\mathbb{P}_{X_0}(x) \\
&= \int_{\mathbb{H}} \left(\int_{\mathbb{R}} \Phi(x,y) f_1^x(y) dy \right) d\mathbb{P}_{X_0}(x) \\
&= \int_{\mathbb{H}\times\mathbb{R}} \Phi(x,y)|D| f_1^x(y) \left(\frac{1}{|D|} \mathbf{1}_D(y) \right) dy d\mathbb{P}_{X_0}(x) \\
&= \int_{\mathbb{H}\times\mathbb{R}} \Phi(x,y)|D| f_1^x(y) d\mathbb{P}_0(x,y).
\end{aligned}$$

Consequently, $\mathbb{P}_1 << \mathbb{P}_0$, and $d\mathbb{P}_1/d\mathbb{P}_0(x,y) = |D| f_1^x(y)$. This enables to compute the Kullback information

$$\begin{aligned}
K(\mathbb{P}_1, \mathbb{P}_0) &= \int \ln\left(\frac{d\mathbb{P}_1}{d\mathbb{P}_0} \right) d\mathbb{P}_1 = \int_{\mathbb{H}\times\mathbb{R}} \ln\left(|D| f_1^x(y) \right) f_1^x(y) dy d\mathbb{P}_{X_0}(x) \\
&= \mathbb{E}\left[\int_D \ln\left(|D| f_1^{X_0}(y) \right) f_1^{X_0}(y) dy \right] \\
&= \mathbb{E}\left[\int_D \ln\left(1 + |D| L\eta_n^\beta H\left(\frac{\|X_0 - x_0\|}{\eta_n} \right) \psi(y) \right) \left(\frac{1}{|D|} + L\eta_n^\beta H\left(\frac{\|X_0 - x_0\|}{\eta_n} \right) \psi(y) \right) dy \right].
\end{aligned}$$

Noting that $\ln(1+u) \leq u$ for every $u > -1$, we obtain

$$\begin{aligned}
K(\mathbb{P}_1, \mathbb{P}_0) &\leq \mathbb{E}\left[\int_D |D| L\eta_n^\beta H\left(\frac{\|X_0 - x_0\|}{\eta_n} \right) \psi(y) dy \right] \\
&\quad + \mathbb{E}\left[\int_D |D| \left(L\eta_n^\beta H\left(\frac{\|X_0 - x_0\|}{\eta_n} \right) \psi(y) \right)^2 dy \right] \\
&= 0 + |D| L^2 \eta_n^{2\beta} \int_D \psi^2(y) dy \mathbb{E}\left[H^2\left(\frac{\|X_0 - x_0\|}{\eta_n} \right) \right] \\
&\leq |D| L^2 \eta_n^{2\beta} \|\psi\|_{L^2(\mathbb{R})}^2 \|H\|_{L^\infty(\mathbb{R})}^2 \mathbb{P}\left(\|X_0 - x_0\| \leq \eta_n \right) \\
&= |D| L^2 \eta_n^{2\beta} \|\psi\|_{L^2(\mathbb{R})}^2 \|H\|_{L^\infty(\mathbb{R})}^2 \varphi^{x_0}(\eta_n),
\end{aligned}$$

by using successively that $\int_{\mathbb{R}} \psi(y)dy = 0$ and that the support of H is $[0;1]$.

Thus, thanks to the definition of η_n, we get $K(\mathbb{P}_1, \mathbb{P}_0) \leq |D| L^2 \|\psi\|_{L^2(\mathbb{R})}^2 \|H\|_{L^\infty(\mathbb{R})}^2 c_{(C)}/n$, and finally,

$$K(\mathbb{P}_1^{\otimes n}, \mathbb{P}_0^{\otimes n}) = nK(\mathbb{P}_1, \mathbb{P}_0) \leq |D| L^2 \|\psi\|_{L^2(\mathbb{R})}^2 \|H\|_{L^\infty(\mathbb{R})}^2 c_{(C)},$$

which completes the proof of (C).

□

Proof of (i), under Assumption $H_{X,M}$ or $H_{X,F}$

The proofs exactly follow the same scheme as for (i) under $H_{X,L}$. The only difference is the choice of the sequence $(\eta_n)_n$ (see (3.41)).

- Under $H_{X,M}$, we set $r_n = \exp\left(-\frac{2\beta}{c_1^{1/\alpha}} \ln^{1/\alpha} n\right)$, and replace the previous choice (3.41) of η_n by $\eta_n = \exp(-(c_1^{-1} \ln n - c_1^{-(\alpha+1)/\alpha}(2\beta+\gamma_2) \ln^{1/\alpha} n)^{1/\alpha})$. It verifies both of the required conditions (3.40).
- The case $H_{X,F}$ is the extreme case $\alpha = 1$ in $H_{X,M}$. We set $\eta_n = n^{1/(2\beta+\gamma)}$ and attain the lower bound $r_n = n^{-2\beta/(2\beta+\gamma)}$.

□

Proof of (ii)

The risk (3.5) is an integral w.r.t the measure $\mathbb{P}_X \otimes \mathbb{P}_{\mathcal{U}_D}$ where $\mathbb{P}_{\mathcal{U}_D}$ is the uniform distribution on the set D. The tools defined to prove (i) are useful and we refer to it. But it cannot be straightforwardly adapted, since for an integrated criterion, two hypotheses are not sufficient. We focus on the case of Assumption $H_{X,L}$ (the switch to Assumption $H_{X,M}$ and $H_{X,F}$ is the same as in (i)). Denote by $r_n = (\ln(n))^{-2\beta/\alpha}$ the rate of convergence again. We must build a set of functions $(F_\omega)_{\omega\in\Omega_n}$ where Ω_n is a non-empty subset of $\{0,1\}^{m_n}$ and m_n is a positive integer which will be precised later, such that:

(A') F_ω belongs to $\mathcal{F}_\beta$, for all $\omega \in \Omega_n$,

(B') for all $\omega, \omega' \in \Omega_n$, $\omega \neq \omega'$, $\mathbb{E}\left[\|F_\omega^{X'} - F_{\omega'}^{X'}\|_D^2 \mathbf{1}_B(X')\right] \geq c r_n$ where $c > 0$ is a constant,

(C') for all $\omega \in \Omega_n$, $\mathbb{P}_\omega$ is absolutely continuous with respect to $\mathbb{P}_0$ and

$$\frac{1}{\mathrm{Card}(\Omega_n)} \sum_{\omega\in\Omega_n} K(\mathbb{P}_\omega^{\otimes n}, \mathbb{P}_0^{\otimes n}) \leq \zeta \ln(\mathrm{Card}(\Omega_n))$$

for a real number $\zeta \in]0, 1/8[$, where $\mathbb{P}_\omega^{\otimes n}$ is the probability distribution of a sample $(X_{\omega,i}, Y_{\omega,i})_{i=1,\dots,n}$ for which the conditional c.d.f. of $Y_{\omega,i}$ given $X_{\omega,i}$ is given by F_ω.

We follow the same steps as previously : first we show that the proof is complete if we are able to find a set of functions $(F_\omega)_{\omega\in\Omega_n}$ verifying hypotheses (A'), (B') and (C') and second we explicit the construction of such a set.

Suppose that the family $(F_\omega)_{\omega\in\Omega_n}$ verifies (A'), (B') and (C'). With arguments and notations similar than the ones of the proof of (i) (p.115), we obtain

$$\inf_{\widehat{F}} \sup_{F\in\mathcal{F}_\beta} \mathbb{E}_F\left[\|\widehat{F}^{X'} - F^{X'}\|_D^2\right] \geq \frac{c}{2} r_n \inf_{\widehat{F}} \sup_{F\in\mathcal{F}_\beta} \mathbb{P}_F\left(\int_B \|\widehat{F}^x - F_{\omega'}^x\|_D^2 dP_X(x) \geq \frac{c}{2} r_n\right),$$

where $\mathbb{E}_F$ (resp. $\mathbb{P}_F$) denotes the expectation (resp. the probability) with respect to the law of $\{(X_i, Y_i), i = 1, \dots, n, X'\}$ when, for all $i = 1, \dots, n$, for all $x \in \mathbb{H}$, the conditional

cumulative distribution of Y_i given $X_i = x$ is F^x and the second infimum is taken over all tests with $\mathrm{Card}(\Omega_n)$ hypotheses calculated from the sample $\{(X_i, Y_i), i = 1, \ldots, n\}$. Then the result comes from the following theorem which can be deduced from Tsybakov (2009, Theorem 2.5, p.99).

Theorem 8. *Suppose that there exists a set Ω_n such that $Card(\Omega_n) \geq 2$ and a family $(F_\omega)_{\omega\in\Omega_n}$ of elements of $\mathcal{F}_\beta$ such that, for a real number $s > 0$,*

- *For all $\omega, \omega' \in \Omega_n$, $\omega \neq \omega'$,*

$$\mathbb{E}\left[\|F_\omega^{X'} - F_{\omega'}^{X'}\|_D^2 \mathbf{1}_B(X')\right] \geq 2s.$$

- *For all $\omega \in \Omega_n$, $\mathbb{P}_\omega$ is absolutely continuous with respect to $\mathbb{P}_0$ and there exists a real number $\zeta \in]0, 1/8[$ such that*

$$\frac{1}{\mathrm{Card}(\Omega_n)} \sum_{\omega\in\Omega_n} K(\mathbb{P}_\omega^{\otimes n}, \mathbb{P}_0^{\otimes n}) \leq \zeta \ln(\mathrm{Card}(\Omega_n)).$$

Then

$$\inf_{\widehat{F}} \sup_{F\in\mathcal{F}_\beta} \mathbb{P}_F\left(\int_B \|\widehat{F}^x - F_{\omega'}^x\|_D^2 dP_X(x) \geq s\right)$$
$$\geq \frac{\sqrt{\mathrm{Card}(\Omega_n)}}{1+\sqrt{\mathrm{Card}(\Omega_n)}}\left(1 - 2\zeta - \sqrt{\frac{2\zeta}{\ln(\mathrm{Card}(\Omega_n))}}\right) \geq c' > 0,$$

with $c' = \frac{\sqrt{2}}{1-\sqrt{2}}(7/8 - 1/(2\ln 2))$.

Construction of the set of hypotheses F_ω and of the associated samples

The first function $(x, y) \mapsto F_0^x(y)$ is defined as in the proof of (i). For all $\omega = (\omega_1, \ldots, \omega_{m_n}) \in \{0, 1\}^{m_n}$, let

$$F_\omega^x(y) = F_0^x(y) + L\eta_n^\beta \int_{-\infty}^y \psi(t)dt \sum_{k=1}^{m_n} \omega_k H\left(\frac{\|x - x_k\|}{\eta_n}\right),$$

where H, L, and $(\eta_n)_n$ are introduced in the proof of (i) (a good choice of η_n is (3.41)), $\psi : \mathbb{R} \to \mathbb{R}$ is a non-zero continuous function with support D such that $\int_{\mathbb{R}} \psi(t)dt = 0$ and $\|\psi\|_{\mathbb{L}^2(\mathbb{R})}^2 < \ln(2)/(64C_2|D|\|H\|_{\mathbb{L}^2(\mathbb{R})}c_{(C)})$ (where C_2 appears in Assumption (3.13), p.85 and $c_{(C)}$ in (3.41)) and with $x_j = \sqrt{2}\sup_{n\in\mathbb{N}^*}\{\eta_n\}e_j$, for all $j \geq 1$, where $(e_j)_{j\geq 1}$ is an orthonormal basis of $\mathbb{L}^2([0, 1])$.

We also choose a variable Y_ω, such that, for any $x \in \mathbb{H}$, the conditional distribution of Y_ω given $X = x$ is characterized by the c.d.f. F_ω^x. The notation $\mathbb{P}_\omega$ is the distribution of (X, Y_ω).

Remark that the definition of $(x_j)_{j=1,\ldots,m_n}$ implies that,

$$H\left(\frac{\|x - x_k\|}{\eta_n}\right) H\left(\frac{\|x - x_j\|}{\eta_n}\right) = 0 \text{ for all } x \in \mathbb{H}, \text{ as soon as } j \neq k. \tag{3.42}$$

Indeed, suppose that $H(\|x-x_k\|/\eta_n) \neq 0$, since H is supported on $[0,1]$, we have $\|x-x_k\| \leq \eta_n$. Now remark that, as $(e_j)_{j\geq 1}$ is an orthonormal basis, for all $j \neq k$, $\|x_j - x_k\|^2 = 2\sup_{n\in\mathbb{N}^*}\{\eta_n^2\}\left(\|e_j\|^2 - 2\langle e_j, e_k\rangle + \|e_k\|^2\right) = 4\sup_{n\in\mathbb{N}^*}\{\eta_n^2\}$. Then $\|x - x_j\| \geq \|x_j - x_k\| - \|x - x_k\| \geq 2\sup_{n\in\mathbb{N}^*}\{\eta_n\} - \eta_n > \eta_n$ and $H(\|x - x_k\|/\eta_n) = 0$.

Checks of the conditions (A') to (C')

Check (A') We have already checked that F_0 belongs to $\mathcal{F}_\beta$. Let $\omega \in \{0,1\}^{m_n}$ be fixed. To prove that F_ω^x is non increasing ($x \in \mathbb{H}$ fixed), as for F_1^x, we bound,

$$\begin{aligned}(F_\omega^x)'(y) &\geq \frac{1}{|D|} - L\eta_n^\beta \|H\|_{L^\infty(\mathbb{R})}\|\psi\|_{L^\infty(\mathbb{R})} \\ &\geq \frac{1}{|D|} - L\sup_{n\in\mathbb{N}^*}\{\eta_n^\beta\}\|H\|_{L^\infty(\mathbb{R})}\|\psi\|_{L^\infty(\mathbb{R})} > 0,\end{aligned}$$

for $y \in D$, thanks to Property (3.42) and the definition of L above. Thus, as F_1 in the proof of (i), F_ω is a conditional distribution function, and we also similarly obtain $F_\omega \in \mathcal{F}_\beta$.

Check (B') For all $\omega, \omega' \in \{0,1\}^{m_n}$,

$$\mathbb{E}\left[\|F_\omega^{X'} - F_{\omega'}^{X'}\|_D^2 \mathbf{1}_B(X')\right] = L^2\eta_n^{2\beta} I_\psi \mathbb{E}\left[\left(\sum_{k=1}^{m_n}(\omega_k - \omega_k')H\left(\frac{\|X' - x_k\|}{\eta_n}\right)\right)^2 \mathbf{1}_B(X')\right],$$

with I_ψ defined in the proof of (i). From Property (3.42), we get:

$$\mathbb{E}\left[\|F_\omega^{X'} - F_{\omega'}^{X'}\|_D^2 \mathbf{1}_B(X')\right] = L^2\eta_n^{2\beta} I_\psi \sum_{k=1}^{m_n}(\omega_k - \omega_k')^2 \mathbb{E}\left[H^2\left(\frac{\|X' - x_k\|}{\eta_n}\right)\mathbf{1}_B(X')\right].$$

Now set $c_H := \min_{x,\|x\|\leq 1/2} H(x)$, since H is continuous and $H(x) > 0$ for all $x \in \mathbb{H}$, such that $\|x\| \leq 1$, we have $c_H > 0$ and

$$\begin{aligned}\mathbb{E}\left[H^2\left(\frac{\|X' - x_k\|}{\eta_n}\right)\mathbf{1}_B(X')\right] &\geq \mathbb{E}\left[H^2\left(\frac{\|X' - x_k\|}{\eta_n}\right)\mathbf{1}_{\left\{\frac{\|X'-x_k\|}{\eta_n}\leq 1/2\right\}}\mathbf{1}_B(X')\right] \\ &\geq c_H^2 \mathbb{P}\left(\left\{\|X' - x_k\| \leq \eta_n/2\right\} \cap \left\{X' \in B\right\}\right).\end{aligned}$$

Now recall that, by definition, $\|x_k\| = \sqrt{2}\sup_{n\in\mathbb{N}}\{\eta_n\}$, and that B contains the ball of $\mathbb{H}$ centred at 0 and of radius ρ. Then, as soon as, $\rho > (1/2 + \sqrt{2})\sup_{n\in\mathbb{N}}\{\eta_n\}$, we have $\{\|X' - x_k\| \leq \eta_n/2\} \subset \{\|X'\| \leq \rho\} \subset \{X' \in B\}$. Then, since $\|x_k\| < \rho$, we also have $x_k \in B$ and we can apply Condition (3.13) to get a lower bound on the shifted small ball probability $\mathbb{P}\left(\|X' - x_k\| \leq \eta_n/2\right) = \varphi^{x_k}(\eta_n/2)$. We get

$$\mathbb{E}\left[H^2\left(\frac{\|X' - x_k\|}{\eta_n}\right)\mathbf{1}_B(X')\right] \geq c_H^2 c_2 \varphi(\eta_n/2),$$

and

$$\mathbb{E}\left[\|F_\omega^{X'} - F_{\omega'}^{X'}\|_D^2 \mathbf{1}_B(X')\right] \geq L^2 c_H^2 c_2 \eta_n^{2\beta} I_\psi \varphi(\eta_n/2) r(\omega, \omega'),$$

where r is the Hamming distance on $\{0,1\}^{m_n}$ defined by $r(\omega,\omega') = \sum_{j=1}^{m_n} \mathbf{1}_{\{\omega_k \neq \omega_k\}}$. Now, from Varshamov-Gilbert bound (Lemma 2.7 of Tsybakov 2009), there exists a subset Ω_n of $\{0,1\}^{m_n}$ such that

$$r(\omega,\omega') \geq \frac{m_n}{8}, \text{ for all } \omega,\omega' \in \Omega_n, \omega \neq \omega', \text{ and } \mathrm{Card}(\Omega_n) \geq 2^{m_n/8}. \tag{3.43}$$

Then fix $m_n := \lfloor \varphi(\eta_n/2)^{-1} \rfloor$ where $\lfloor \cdot \rfloor$ is the integer part. For all $\omega \neq \omega'$, by definition of η_n

$$\mathbb{E}\left[\|F_\omega^{X'} - F_{\omega'}^{X'}\|_D^2 \mathbf{1}_B(X') \right] \geq \frac{1}{8} L^2 c_H^2 c_2 \eta_n^{2\beta} I_\psi m_n \varphi(\eta_n/2) \geq \frac{1}{8} L^2 c_H^2 c_2 r_n.$$

Check (C') We also prove that the measure $\mathbb{P}_\omega$ is absolutely continuous with respect to $\mathbb{P}_0$, with derivative $d\mathbb{P}_\omega / d\mathbb{P}_0(x,y) = |D| f_\omega^x(y)$ and $d\mathbb{P}_\omega(x,y) = f_\omega^x(y) dy d\mathbb{P}_X(x)$, like in the proof of (i). Arguing again as in (i), we get

$$\begin{aligned} K(\mathbb{P}_\omega, \mathbb{P}_0) &\leq |D| L^2 \eta_n^{2\beta} \int_D \psi^2(y) dy \sum_{k=1}^{m_n} \omega_k \mathbb{E}\left[H^2\left(\frac{\|X - x_k\|}{\eta_n} \right) \right] \\ &\leq m_n |D| L^2 \eta_n^{2\beta} \|\psi\|_{L^2(\mathbb{R})}^2 \|H\|_{L^\infty(\mathbb{R})}^2 \mathbb{P}\left(\|X - x_k\| \leq \eta_n \right). \end{aligned}$$

Now, arguing again as in Check (A'), we can apply Assumption (3.13) and get that $\mathbb{P}(\|X - x_k\| \leq \eta_n) \leq C_2 \mathbb{P}(\|X\| \leq \eta_n) = C_2 \varphi(\eta_n)$. Thanks to the definition of η_n, we now obtain (as in (i))

$$K(\mathbb{P}_\omega^{\otimes n}, \mathbb{P}_0^{\otimes n}) \leq C_2 m_n |D| L^2 \|\psi\|_{L^2(\mathbb{R})}^2 \|H\|_{L^\infty(\mathbb{R})}^2 c_{(C)}.$$

Finally, condition (3.43) on the cardinal of Ω_n leads to $m_n \leq (8/\ln 2) \ln(\mathrm{Card}(\Omega_n))$, which completes the proof of (C'), and at the same time the proof of all the lower bounds.

□

3.7.7 Proof of Proposition 2

Main part of the proof

The proof starts like the proof of Theorem 3. For Inequality (ii), we first bound $\mathbb{E}[\|\widehat{F}_h^{X'} - F^{X'}\|_D^2 \mathbf{1}_{\{R_h^{X'} < 1/2\}} \mathbf{1}_B(X')]$ by $C/(n\varphi_p(h))$, with d replaced by d_p in the definition of $R_h^{X'}$. For $\mathbb{E}[\|\widehat{F}_h^{X'} - F^{X'}\|_D^2 \mathbf{1}_{\{R_h^{X'} \geq 1/2\}} \mathbf{1}_B(X')]$ we obtain the splitting (3.22). Lemmas 9 and Lemmas 10 remain valid (by replacing again d by d_p in every terms, and by using H'_φ instead of H_φ). This first part is also easily adapted to the proof of Inequality (i).

The difference lies in the control of the bias term. We substitute to Lemma 11 the following result, the proof of which can be found below. This ends the proof.

Lemma 14. *Suppose that Assumptions H_F and H_ξ are fulfilled. Then*

$$\mathbb{E}\left[\left\| F^{X'} - \mathbb{E}_{X'}\left[\widetilde{F}_{h,p}^{X'} \right] \right\|_D^2 \right] \leq C \left(h^{2\beta} + \left(\sum_{j>p} \sigma_j^2 \right)^\beta \right)$$

and

$$\mathbb{E}\left[\left\|F^{x_0}-\mathbb{E}\left[\widetilde{F}^{x_0}_{h,p}\right]\right\|^2_D\right]\leq C\left(h^{2\beta}+\left(\sum_{j>p}\sigma_j^2\right)^{\beta}+\left(\sum_{j>p}\langle x_0,e_j\rangle^2\right)^{\beta}\right),$$

where $C>0$ only depends on C_D, β, and C_ξ.

□

Proof of Lemma 14

Let us begin with the first inequality (integrated risk). Like in the proof of Lemma 11, we also obtain (3.23). Then,

$$\mathbb{E}\left[\left\|F^{X'}-\mathbb{E}_{X'}\left[\widetilde{F}^{X'}_{h,p}\right]\right\|^2_D\right] \leq C_D^2n^2\mathbb{E}\left[\mathbb{E}_{X'}\left[\widetilde{W}_h^{(1)}(X')\left(d_p^2(X_1,X')+\sum_{j>p}\sigma_j^2(\xi_j^{(1)}-\xi_j')^2\right)^{\beta/2}\right]^2\right],$$

where $\xi_j^{(1)}:=(\langle X_1,e_j\rangle-\mu_j)/\sigma_j$ and $\xi_j':=(\langle X_1,e_j\rangle-\mu_j)/\sigma_j$ are the standardized versions of $\langle X_1,e_j\rangle$ and $\langle X',e_j\rangle$. The same arguments as in Lemma 11 lead to

$$\mathbb{E}\left[\left\|F^{X'}-\mathbb{E}_{X'}\left[\widetilde{F}^{X'}_{h,p}\right]\right\|^2_D\right]\leq C_D^2n^2\mathbb{E}\left[\mathbb{E}_{X'}\left[\widetilde{W}_h^{(1)}(X')\left(h^2+\sum_{j>p}\sigma_j^2(\xi_j^{(1)}-\xi_j')^2\right)^{\beta/2}\right]^2\right].$$

Now, firstly, for all $a,b>0$, $(a+b)^{\beta/2}\leq(2\max\{a,b\})^{\beta/2}\leq 2^{\beta/2}(a^{\beta/2}+b^{\beta/2})$ and secondly $\mathbb{E}_{X'}\left[\widetilde{W}_h^{(1)}(X')\right]=1/n$. We thus obtain

$$\mathbb{E}\left[\left\|F^{X'}-\mathbb{E}_{X'}\left[\widetilde{F}^{X'}_{h,p}\right]\right\|^2_D\right] \leq 2^{\beta+1}C_D^2\left(h^{2\beta}+n^2\mathbb{E}\left[\mathbb{E}_{X'}\left[\widetilde{W}_h^{(i)}(X')\left(\sum_{j>p}\sigma_j^2(\xi_j^{(i)}-\xi_j')^2\right)^{\beta/2}\right]^2\right]\right). \quad (3.44)$$

Under Assumption H_ξ^b, the results come from the following bound

$$\mathbb{E}\left[\mathbb{E}_{X'}\left[\widetilde{W}_h^{(1)}(X')\left(\sum_{j>p}\sigma_j^2(\xi_j^{(1)}-\xi_j')^2\right)^{\beta/2}\right]^2\right]\leq 4^\beta C_\xi^{2\beta}\left(\sum_{j>p}\sigma_j^2\right)^{\beta}\frac{1}{n^2}.$$

Under Assumption H_ξ^{ind}, remark that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_{X'}\left[\widetilde{W}_h^{(1)}(X')\left(\sum_{j>p}\sigma_j^2(\xi_j^{(1)}-\xi_j')^2\right)^{\beta/2}\right]^2\right] \\
&\qquad = \mathbb{E}\left[\mathbb{E}_{X'}\left[\widetilde{W}_h^{(1)}(X')\right]^2\mathbb{E}_{X'}\left[\left(\sum_{j>p}\sigma_j^2(\xi_j^{(1)}-\xi_j')^2\right)^{\beta/2}\right]^2\right] \\
&\qquad \leq n^{-2}\mathbb{E}\left[\left(\sum_{j>p}\sigma_j^2(\xi_j^{(1)}-\xi_j')^2\right)^{\beta/2}\right]^2.
\end{aligned}$$

Now applying Lemma 15 below with $\eta_j = \xi_j^{(1)} - \xi_j'$ and $C_M = 2^\beta C_\xi$, we get

$$\mathbb{E}\left[\left(\sum_{j>p}\sigma_j^2(\xi_j^{(1)}-\xi_j')^2\right)^{\beta/2}\right]^2 \leq 2^{4\beta}C_\xi^2\left(\sum_{j>p}\sigma_j^2\right)^\beta,$$

and the result comes from Inequality (3.44). The proof of the first inequality of Lemma 14 is completed.

For the second inequality (pointwise risk), the only difference is that, from (3.23), we rather use

$$\begin{aligned}
\|X_1 - x_0\|^\beta &= \left(d_p^2(X_1,x_0) + \sum_{j>p}(\sigma_j\xi_j - \langle x_0, e_j\rangle)^2\right)^{\beta/2} \\
&\leq 3^{\beta/2}\left(d_p^\beta(X_1,x_0) + 2^{\beta/2}\left(\sum_{j>p}\sigma_j^2\xi_j^2\right)^{\beta/2} + 2^{\beta/2}\left(\sum_{j>p}\langle x_0,e_j\rangle^2\right)^{\beta/2}\right).
\end{aligned}$$

The final bound then follows similarly.

□

Lemma 15. *Let $(\eta_j)_{j\geq 1}$ a sequence of real random variables and $(\sigma_j)_{j\geq 1}$ a sequence a real numbers verifying, for $\beta > 0$,*

$$\sum_{j\geq 1}\sigma_j^2 < +\infty \text{ and } \forall j \geq 1,\ \mathbb{E}\left[\eta_j^\beta\right] \leq C_M,$$

for a constant $C_M > 1$, then, for all $p \in \mathbb{N}$

$$\mathbb{E}\left[\left(\sum_{j>p}\sigma_j^2\eta_j^2\right)^{\beta/2}\right] \leq C_M\left(\sum_{j>p}\sigma_j^2\right)^{\beta/2}.$$

Proof of Lemma 15 First suppose that $\beta/2 \in \mathbb{N}^*$, we have

$$\left(\sum_{j>p}\sigma_j^2\eta_j^2\right)^{\beta/2} = \sum_{j_1,\ldots,j_{\beta/2}>p}\prod_{l=1}^{\beta/2}\sigma_{j_l}^2\eta_{j_l}^2,$$

and, by a classical generalization of Hölder's Inequality

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{j>p}\sigma_j^2\eta_j^2\right)^{\beta/2}\right] &= \sum_{j_1,\ldots,j_{\beta/2}>p}\prod_{l=1}^{\beta/2}\sigma_{j_l}^2\mathbb{E}\left[\prod_{l=1}^{\beta/2}\eta_{j_l}^2\right] \leq \sum_{j_1,\ldots,j_{\beta/2}>p}\prod_{l=1}^{\beta/2}\sigma_{j_l}^2\prod_{l=1}^{\beta/2}\mathbb{E}\left[\eta_{j_l}^{\beta}\right]^{2/\beta}\\ &\leq C_M\sum_{j_1,\ldots,j_{\beta/2}>p}\prod_{l=1}^{\beta/2}\sigma_{j_l}^2 \leq C_M\left(\sum_{j>p}\sigma_j^2\right)^{\beta/2}.\end{aligned}$$

Now suppose that $\beta \in \mathbb{Q}\cap]0,+\infty[$, we can write without loss of generality that $\beta/2 = p/q$ with $p \in \mathbb{N}^*$ and $q > 1$ (if $q = 1$, $\beta/2 \in \mathbb{N}^*$). Then the function $x \mapsto x^{1/q}$ is concave and by Jensen's Inequality:

$$\mathbb{E}\left[\left(\sum_{j>p}\sigma_j^2\eta_j^2\right)^{\beta/2}\right] \leq \mathbb{E}\left[\left(\sum_{j>p}\sigma_j^2\eta_j^2\right)^{p}\right]^{1/q} \leq C_M^{1/q}\left(\sum_{j>p}\sigma_j^2\right)^{\beta/2} \leq C_M\left(\sum_{j>p}\sigma_j^2\right)^{\beta/2}.$$

The case $\beta > 0$ follows immediately from the fact that $\mathbb{Q}$ is dense in $\mathbb{R}$.

□

3.7.8 Proof of Corollary 1

The proof is based on the same ideas as the ones used to prove Theorem 5 in Section 3.7.5. We begin with the result (ii) (integrated risk).

Proof of (ii)

Thanks to Proposition 2 (ii), the risk of the estimator is bounded by $h^{2\beta} + (\sum_{j>p}\sigma_j^2)^{\beta} + n^{-1}\varphi_p^{-1}(h)$, up to a multiplicative constant. Remark that

$$\varphi_p(h) = \int_{\left\{\mathbf{x}\in\mathbb{R}^p,\ \sum_{j=1}^p\sigma_j^2x_j^2\leq h^2\right\}} f_p(x_1,\ldots,x_p)d\mathbf{x},$$

where f_p is the density of $(\xi_1, \ldots, \xi_p)$. By noticing that

$$\left\{ \mathbf{x} \in \mathbb{R}^p,\ \sum_{j=1}^{p} \sigma_j^2 x_j^2 \leq h^2 \right\} \supset \left\{ \mathbf{x} \in \mathbb{R}^p,\ |x_j| \leq \frac{h}{\sqrt{\sum_{j=1}^{p} \sigma_j^2}} \right\},$$

we get

$$\varphi_p(h) \ \geq \ 2^p h^p \int_{\left[0, \left(\sum_{j=1}^{p} \sigma_j^2\right)^{-1/2}\right]^p} f_p(hx_1, \ldots, hx_p) d\mathbf{x} \geq c h^p,$$

where c only depends on $\sum_{j\geq 1} \sigma_j^2$ and $f_p(0)$ and p. With the assumption on σ_j, we thus obtain the following upper bound for the risk, up to a constant $R(h,p) := h^{2\beta} + p^{\beta(1-2\delta)} + c^{-p} n^{-1} h^{-p}$. We then compute the partial derivatives

$$\begin{aligned} \frac{\partial R}{\partial h}(h,p) &= 2\beta h^{2\beta-1} - p c^{-p} n^{-1} h^{-p-1}, \\ \frac{\partial R}{\partial p}(h,p) &= \beta(1-2\delta) p^{\beta(1-2a)-1} - \ln(ch) c^{-p} n^{-1} h^{-p}. \end{aligned}$$

If (h^*, p^*) is the minimizer of R we have $\frac{\partial R}{\partial h}(h^*, p^*) = 0$, which leads to

$$h^* = \left(\frac{p^* c^{-p^*}}{2\beta} \right)^{1/(2\beta+p^*)} n^{-1/(2\beta+p^*)}.$$

Moreover, for all $p \in \mathbb{N}^*$,

$$\begin{aligned} R(h^*, p^*) &= \left(\frac{p^* c^{-p^*}}{2\beta} \right)^{2\beta/(2\beta+p^*)} n^{-2\beta/(2\beta+p^*)} + (p^*)^{\beta(1-2\delta)} \\ &\quad + c^{-p} n^{-1} \left(\frac{p^* c^{-p^*}}{2\beta} \right)^{-p^*/(2\beta+p^*)} n^{-p^*/(2\beta+p^*)} \\ &\leq \left(\frac{p c^{-p}}{2\beta} \right)^{2\beta/(2\beta+p)} n^{-2\beta/(2\beta+p)} + p^{\beta(1-2\delta)} \\ &\quad + c^{-p} n^{-1} \left(\frac{p c^{-p}}{2\beta} \right)^{-p/(2\beta+p)} n^{-p/(2\beta+p)}, \end{aligned}$$

and this last bound has the order $n^{-2\beta/(2\beta+p)} + p^{\beta(1-2\delta)}$. Choosing $h = (pc^{-p}/2\beta)^{1/(2\beta+p)} n^{-1/(2\beta+p)}$ and $p = [\ln(n)/(\delta - 1/2)\ln\ln(n) - 2\beta]$ gives the result.

Proof of (i)

We deduce from Proposition 2 (ii), from the assumption $\sum_{j>p} \langle x_0, e_j \rangle \leq C \sum_{j>p} \sigma_j^2$, and from the left-hand-side inequality of H_φ that the risk is upper bounded by $h^{2\beta} + (\sum_{j>p} \sigma_j^2)^\beta + n^{-1} \varphi_p^{-1}(h)$, up to a multiplicative constant. Thus, the reasoning is the same as for (ii).

□

Chapitre 4

Méthodologie des surfaces de réponse et données fonctionnelles

L'objectif de ce chapitre est d'adapter au cadre fonctionnel la méthodologie des surfaces de réponse qui consiste, dans sa version classique, à trouver les valeurs des covariables $x_1, \ldots, x_d$, qui sont des paramètres d'une certaine expérience, pour laquelle la valeur de la réponse Y (le résultat de l'expérience) est optimale. Cette optimisation se fait en réalisant des expériences supplémentaires qui peuvent être coûteuses et doivent donc être limitées en nombre. Nous présentons dans un premier temps la version classique, multivariée, de la méthodologie des surfaces de réponse. Nous motivons et proposons ensuite une extension de la méthode des surfaces de réponse pour l'optimisation d'une variable réponse dépendant d'une covariable fonctionnelle. Cette nouvelle méthode est basée principalement sur la génération de plans d'expériences dans un espace fonctionnel. Une étude numérique des méthodes proposées est ensuite présentée incluant une application à des données réelles.

Sommaire

4.1 Introduction

Response Surface Methodology (RSM) was introduced by Box and Wilson (1951) with the goal of identifying optimal conditions for experiments in chemistry. The target was to

minimize the cost of experimentation or maximize the purity of the product obtained by finding the right combination of factors (temperature, pressure, proportion of reactants, ...). Then, its purpose is to find the values of explanatory variables $(x_1, \ldots, x_d) \in \mathbb{R}^d$ for which the response variable is optimal $Y \in \mathbb{R}$. This method has been and is still widely used in the industry.

4.1.1 Response Surface Methodology for multivariate covariate

We first present the guidelines of RSM in its natural context which is the optimization of a real response depending on $d \geq 2$ real covariates.

4.1.1.1 Sequential Optimization

Suppose we want to find the values of $(x_1, \ldots, x_d)' \in \mathcal{R}$ – where $\mathcal{R}$ is a given region of $\mathbb{R}^d$ – minimizing an unknown function $m : \mathcal{R} \to \mathbb{R}$. We assume here that, for all $(x_1, \ldots, x_d)' \in \mathcal{R}$ we observe $Y(x_1, \ldots, x_d) =: Y$ such that $\mathbb{E}[Y|x_1, \ldots, x_n] = m(x_1, \ldots, x_d)$.

The principle of the method is to find the optimal experimental conditions by performing a limited number of experiments. We choose a starting point $\mathbf{x}_0^{(0)} \in \mathcal{R}$ (existing conditions of experimentation that we try to optimize for example) and perform a series of experiments around this point. The obtained observations allow us to have some idea of the shape of the surface of $y = m(x_1, \ldots, x_n)$ in this region. This knowledge is used to estimate the direction of the steepest descent of the surface. Along this direction, a series of experiments are carried out until a point where the response is considered optimal is reached. Another series of experiments may then be realized around this point and the results give another direction of descent, or can be used to refine the position of the optimum.

More precisely, we start the algorithm by choosing a set of points, called *design points*, $\left(\mathbf{x}_i^{(0)}\right)_{i=1,\ldots,n_0} := \left(x_{i,1}^{(0)}, \ldots, x_{i,d}^{(0)}\right)_{i=1,\ldots,n_0}$ around the initial point $\mathbf{x}_0^{(0)}$ and we obtain the corresponding responses $\left(Y_i^{(0)}\right)_{i=1,\ldots,n_0}$.

The obtained data $\left(\left(x_i^{(0)}, Y_i^{(0)}\right), i = 1, \ldots, n_0\right)$ are fitted, typically with a polynomial model of the type

$$Y := \sum_{\substack{0 \leq j_1 \ldots, j_d \leq p \\ j_1 + \ldots + j_d = p}} \beta_{j_1 \ldots, j_d} x_1^{j_1} \ldots x_d^{j_d} + \varepsilon. \tag{4.1}$$

Since the points of the design are supposed to be fairly close to the point $\mathbf{x}_0^{(0)}$, a justification of using this model comes from the Taylor expansion of the function m around $\mathbf{x}_0^{(0)}$ given by

$$\begin{aligned} m(\mathbf{x}) &= m\left(\mathbf{x}_0^{(0)}\right) + \sum_{j=1}^{d} \frac{\partial m}{\partial x_j}\left(\mathbf{x}_0^{(0)}\right)\left(x_j - x_{0,j}^{(0)}\right) + \ldots \\ &+ \sum_{1 \leq j_1, \ldots, j_q \leq d} \frac{\partial^p m}{\partial x_{j_1} \ldots \partial x_{j_p}}\left(\mathbf{x}_0^{(0)}\right)\left(x_{j_1} - x_{0,j_1}^{(0)}\right) \ldots \left(x_{j_p} - x_{0,j_p}^{(0)}\right) + o\left(\left\|\mathbf{x} - \mathbf{x}_0^{(0)}\right\|^p\right), \end{aligned}$$

for all $\mathbf{x} = (x_1, \ldots, x_d)'$ in the interior of $\mathcal{R}$ (here we suppose that m is p-times differentiable).

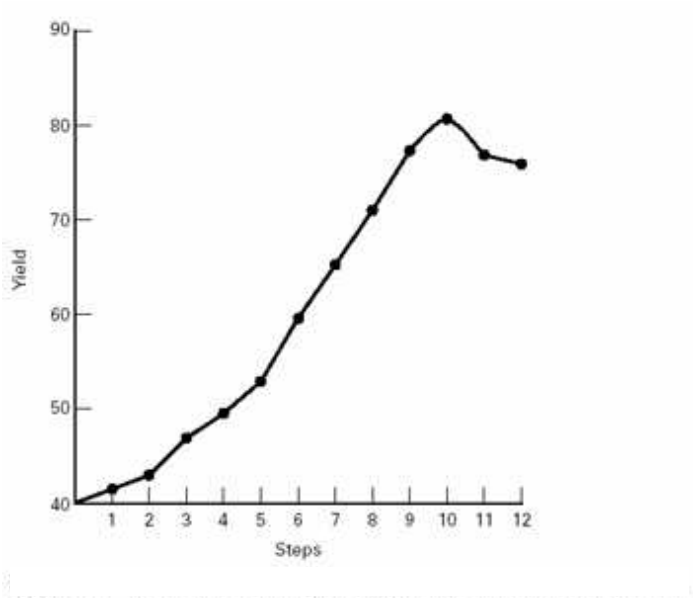


Figure 4.1: Evolution of yield in the direction of steepest descent (Montgomery, 2009).

In particular, remark that $(\beta_{1,0\ldots,0}, \ldots, \beta_{0,\ldots,0,1})'$ is the gradient $\left(\frac{\partial m}{\partial x_1}\left(\mathbf{x}_0^{(0)}\right), \ldots, \frac{\partial m}{\partial x_d}\left(\mathbf{x}_0^{(0)}\right)\right)'$ of m at the point $\mathbf{x}_0^{(0)}$. Note that the usual assumptions on the noise $\varepsilon_1, \ldots, \varepsilon_n$, in particular the assumption of homoscedasticity, are not necessarily verified. However, if the design and the degree p of the model are well chosen, we can suppose that the influence of $\mathbf{x}_i^{(0)}$ in the term $o\left(\left\|\mathbf{x} - \mathbf{x}_0^{(0)}\right\|^p\right)$ is negligible and that the noise is homoscedastic. We also suppose that $\varepsilon_1, \ldots, \varepsilon_n$ are independent and normally distributed.

Once an appropriate model is chosen, least-squares estimators $\widehat{\beta}_{j_1,\ldots,j_d}$ of the coefficients $\beta_{j_1,\ldots,j_d}$ in the model (4.1) are calculated, which gives an idea of the shape of the response surface around $\mathbf{x}_0^{(0)}$.

A second-step is to carry out a series of experiments in the direction of the steepest descent $-(\widehat{\beta}_{1,0\ldots,0}, \ldots, \widehat{\beta}_{0,\ldots,0,1})$ until the response begins to rise. By way of illustration, the following figure shows the evolution of performance along the path of the steepest ascent (here we want to maximize the yield of a chemical reaction as a function of reaction time and temperature). The maximum yield seems to be reached at the tenth step.

At the end of this step, we obtain a new point $\mathbf{x}_0^{(1)}$. The experimenter may decide either to generate new experimental conditions $\left(\mathbf{x}_i^{(1)}\right)_{i=1,\ldots,n_1}$ around $\mathbf{x}_0^{(1)}$ and define another steepest descent direction or to stop and possibly refine optimization using a model of highest degree (typically quadratic $p = 2$).

4.1.1.2 Experimental design in Response Surface Methodology

A key issue in Response Surface Methodology is the choice of the *experimental design* at each step, that is to say the sequences $\{\mathbf{x}_i^{(k)}, i = 1\ldots, n_k\}$. The aim is to choose it so as to get as close as possible to the surface $y = m(x_1, \ldots, x_d)$ while minimizing the number of experiments to perform. This question is still an open problem, we refer to Georgiou, Stylianou, and Aggarwal (2014) and references therein for a complete list of methods used. We will focus here on the most traditional experimental designs.

First-order designs. First-order designs are the experimental designs traditionally used when the considered model is a linear model ((4.1) with $p = 1$). The most frequently employed are factorial designs.

The 2^d *factorial design* is the simplest. For each explanatory variable $x_1, \ldots, x_d$, we choose two levels (coded by $+1$ and -1) and we take all the 2^d combinations of these two levels (see Figure 4.2).

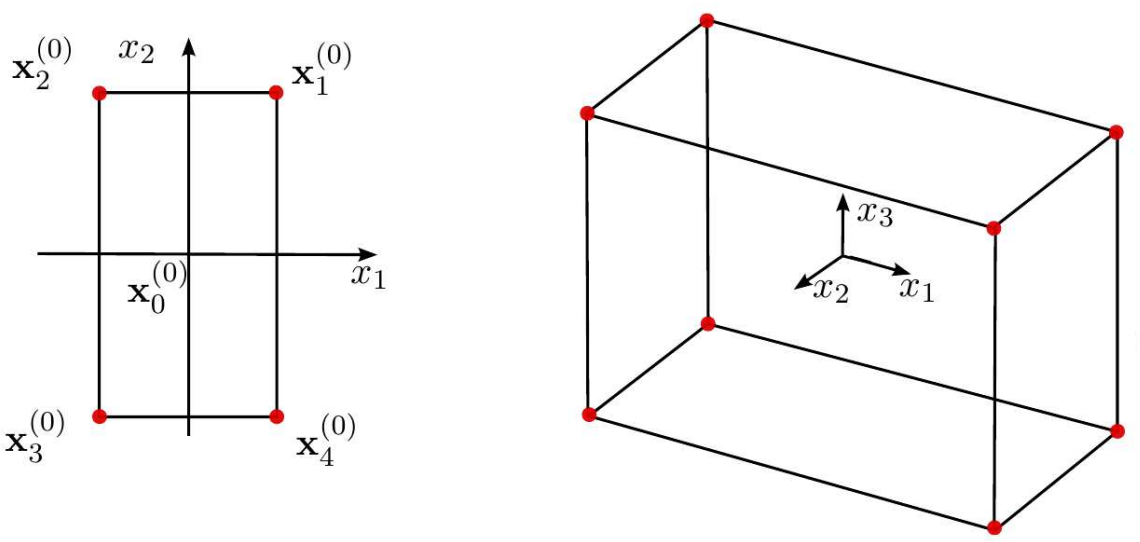


Figure 4.2: 2^d factorial designs for $d = 2$ and $d = 3$.

If d is large, it may be impossible to achieve the 2^d factorial experiments. The *fractional factorial design* keeps only a certain proportion (e.g. a half, a quarter, ...) of points of a 2^d factorial design. Typically, when a fraction $1/(2^p)$ is kept from the original 2^d design, this design is called 2^{d-p} factorial design. The points removed are carefully chosen, we refer e.g. to Gunst and Mason 2009 for more details.

Second-order designs. Second-order designs are those used when it is desired to approach the surface using a quadratic model ((4.1) with $p = 2$). As the number of parameters to estimate is larger, it is necessary to generate more points. Traditional second-order designs are factorial designs, central composite designs and Box-Behnken designs.

- 3^d or 3^{d-p} factorial designs are similar to 2^d and 2^{d-p} factorial designs but with three levels ($+1$, -1 and 0).

– *Central Composite Designs (CCD)* are obtained by adding to the two-level factorial design (fractional or not) two points on each axis of the control variables of both sides of the origin and at distance $\alpha > 0$ (see Figure 4.3 for the case $d = 2$ and $d = 3$). In order to verify some additional properties (see next section *Design Properties*), we also add n_0 copies of the origin.

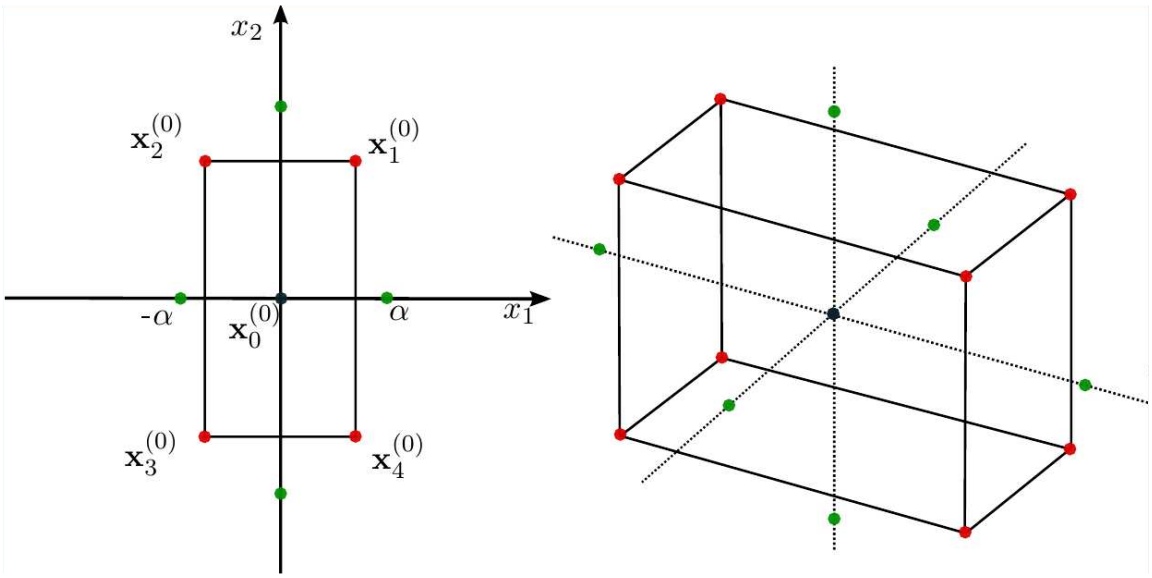


Figure 4.3: Central Composite Designs for $d = 2$ and $d = 3$.

– *Box-Behnken Designs (BBD)* are widely used in the industry. It is a well-chosen subset of the 3^d factorial design. Figure 4.4 represents the construction of Box-Behnken design for $d = 3$. Box-Behnken designs are not used when $d = 2$. For $d \geq 4$, we refer to Myers, Montgomery, and Anderson-Cook (2009, p. 7.4.7).

For all these designs, some or all points may be replicated, this may allow the design to verify some additional properties (see next section) and perform lack-of-fit tests (Brook and Arnold, 1985, pp. 48-49).

4.1.1.3 Design Properties

The major difference appearing here with more traditional statistical contexts is that we can choose the values of the covariates (the design) for which experiments are done. Given the example of polynomial models (4.1), the aim is to choose the design so that the coefficients of the model $(\beta_{j_1,\ldots,j_d})_{\substack{0\leq j_1,\ldots,j_d\leq d\\ j_1+\ldots+j_d=d}}$ are estimated as effectively as possible. There are different ways of conceiving the properties a design should have and therefore there are different criteria used in the literature. We focus on the most classical: orthogonality, rotatability and alphabetic optimality.

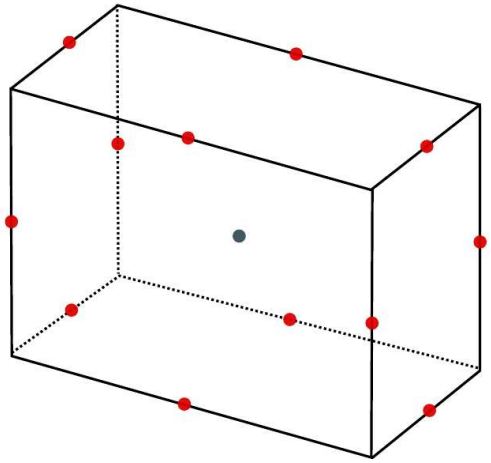


Figure 4.4: Box-Behnken Design for $d = 3$.

Orthogonality. In order to motivate the introduction of the notion of orthogonality, suppose that we are in the k-th step and that the following linear model is verified:

$$Y_i^{(k)} := \sum_{\substack{0\leq j_1\ldots,j_d\leq p\\ j_1+\ldots+j_d=p}} \beta_{j_1\ldots,j_d}^{(k)} \left(x_{i,1}^{(k)}\right)^{j_1} \ldots \left(x_{i,d}^{(k)}\right)^{j_d} + \varepsilon_i^{(k)} \ (i = 1, \ldots, n_k).$$

This model can classically be rewritten in a matrix form

$$Y^{(k)} = \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)} + \boldsymbol{\varepsilon}^{(k)},$$

where $Y^{(k)} = (Y_1^{(k)}, \ldots, Y_{n_k}^{(k)})'$, $\boldsymbol{\beta}^{(k)} = \left(\beta_{j_1\ldots,j_d}^{(k)}\right)_{\substack{0\leq j_1\ldots,j_d\leq d,\\ j_1+\ldots+j_d=d}}$,

$$\mathbf{X}^{(k)} = \left(\left(x_{i,1}^{(k)}\right)^{j_1} \ldots \left(x_{i,d}^{(k)}\right)^{j_d}\right) \ i = 1, \ldots, n_k, \quad \begin{array}{l} 0 \leq j_1 \ldots, j_d \leq p \\ j_1 + \ldots + j_d = p \end{array}$$

and $\boldsymbol{\varepsilon}^{(k)} = (\varepsilon_1^{(k)}, \ldots, \varepsilon_{n_k}^{(k)})'$.

We suppose that $\boldsymbol{\varepsilon}^{(k)}$ is centered, that its components are uncorrelated with same variance σ_k^2. The least-squares estimator of $\beta^{(k)}$ is equal to

$$\widehat{\boldsymbol{\beta}}^{(k)} := \left(\mathbf{X}^{(k)\prime}\mathbf{X}^{(k)}\right)^{-1} \mathbf{X}^{(k)\prime}Y^{(k)}$$

and is a random vector of mean $\mathbb{E}\left[\widehat{\boldsymbol{\beta}}^{(k)}\right] = \boldsymbol{\beta}^{(k)}$ and variance-covariance matrix given by

$$\begin{aligned}\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}^{(k)}\right) &= \mathbb{E}\left[\left(\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}\right)'\right] \\ &= \left(\mathbf{X}^{(k)\prime}\mathbf{X}^{(k)}\right)^{-1}\mathbf{X}^{(k)\prime}\mathbb{E}\left[\boldsymbol{\varepsilon}^{(k)}\boldsymbol{\varepsilon}^{(k)\prime}\right]\mathbf{X}^{(k)}\left(\mathbf{X}^{(k)\prime}\mathbf{X}^{(k)}\right)^{-1} \\ &= \sigma_k^2\left(\mathbf{X}^{(k)\prime}\mathbf{X}^{(k)}\right)^{-1}.\end{aligned}$$

Then, if we can choose the design so that the matrix $\mathbf{X}^{(k)\prime}\mathbf{X}^{(k)}$ is diagonal, the estimated coefficients are uncorrelated. If, in addition, the noise $\boldsymbol{\varepsilon}$ is supposed to be Gaussian, the vector $\widehat{\boldsymbol{\beta}}^{(k)}$ is also a Gaussian random vector with independent components. This makes it easier to test the significance of the components of $\boldsymbol{\beta}^{(k)}$ in the model.

We can verify that 2^d factorial designs are orthogonal first-order designs. However, fractional designs have to be constructed carefully in order to keep the orthogonality property, for instance $\{(1,1),(1,-1)\}$ is a 2^{2-1} design but is not orthogonal. Orthogonality for second-order designs is even harder to verify, we refer to Box and Hunter (1957) for general criteria applied to factorial and fractional factorial designs. Central Composite Designs are orthogonal if

$$\alpha = \sqrt{\frac{\sqrt{F(F+2d+n_0)} - F}{2}},$$

where F is the number of points of the initial factorial design (see Myers, Montgomery, and Anderson-Cook 2009).

Rotatability. To motivate the introduction of rotatability, we investigate the properties of the predicted response at a point $\mathbf{x} = (x_1, \ldots, x_d)' \in \mathbb{R}^d$:

$$\widehat{y}(\mathbf{x})^{(k)} := f_p(\mathbf{x})'\widehat{\boldsymbol{\beta}}^{(k)},$$

where $f_p(\mathbf{x}) := \left(\left(x_1^{(k)}\right)^{j_1}\ldots\left(x_d^{(k)}\right)^{j_d}\right)_{\substack{0\le j_1\ldots,j_d\le p \\ j_1+\ldots+j_d=p}}$. It is easily seen that

$$\mathbb{E}\left[\widehat{y}(\mathbf{x})^{(k)}\right] = f_p(\mathbf{x})'\boldsymbol{\beta}^{(k)}$$

and that

$$\mathrm{Var}(\widehat{y}(\mathbf{x})^{(k)}) = \sigma^2 f_p(\mathbf{x})'\left(\mathbf{X}^{(k)\prime}\mathbf{X}^{(k)}\right)^{-1} f_p(\mathbf{x}).$$

A design is said to be *rotatable* if $\mathrm{Var}(\widehat{y}(\mathbf{x})^{(k)})$ depends only on the distance between $\mathbf{x}$ and the origin. This means that the variance of the response is unchanged under any rotation of the coordinate axes. We refer to Box and Hunter (1957) for conditions of rotatability.

Since $f_1(\mathbf{x}) = (1, x_1, \ldots, x_d)$, we see that all first-order orthogonal designs are also rotatable. This is not the case for second-order designs, for instance a CCD design is rotatable if $\alpha = F^{1/4}$ which means that a CCD design can be rotatable and orthogonal only for some specific values of n_0 and F. Box-Behnken designs are rotatable for $d = 4$

and $d = 7$. Some measures of rotatability have been introduced (Khuri, 1988; Draper and Guttman, 1988; Draper and Pukelsheim, 1990; Park, Lim, and Baba, 1993) in order to compare rotatability of designs.

Alphabetic optimality. This kind of optimality criterion considers some aspects of the variance of estimated model parameter $\widehat{\boldsymbol{\beta}}$. The most important is the *D-optimality criterion* which maximizes the determinant of the matrix $\mathbf{X}^{(k)'}\mathbf{X}^{(k)}$. A justification of such a criterion is to minimize the volume of the confidence region for $\boldsymbol{\beta}$.

Another classical criterion is the *G-optimality criterion* which minimizes the maximal value of $\mathrm{Var}(\widehat{y}(\mathbf{x}))$ over $\mathbf{x} \in \mathcal{R}$.

D-optimal and G-optimal designs may be generated by computers and are used as alternatives to classical designs when they are not available (this is the case for instance when the region $\mathcal{R}$ is constrained).

Other criteria are *A-optimality* minimizing the average variance of the estimated coefficients or *E-optimality* maximizing the minimal eigenvalue of the matrix $\mathbf{X}^{(k)'}\mathbf{X}^{(k)}$. We refer to Pázman (1986) for more details.

4.1.2 Motivating example: temperature, pressure and heat transfer transients in a nuclear reactor vessel

An hypothetical cause of nuclear accident is the loss of coolant accident (LOCA). This is caused by a breach on the primary circuit. In order to avoid reactor meltdown, the safety procedure consists in incorporating cold water in the primary circuit. This can cause a pressurised thermal shock on the nuclear vessel inner wall which increases the risk of failure of the vessel (see Figure 4.5).

The parameters influencing the probability of failure are the evolution over time of temperature, pressure and heat transfer. Obviously, the behavior of the reactor vessel during the accident can be hardly explored by physical experimentation and numerical codes have been developed, for instance by the CEA [1], reproducing the mechanical behavior of the vessel given the three mentioned parameters (temperature, pressure, heat transfer). Figure 4.6 represents different evolution of each parameter during the procedure depending on the value of several input parameters. The colors of each curve depends on the corresponding margin factor (MF) which decreases when the probability of failure of the nuclear reactor increases.

The aim is to find the temperature transient which minimizes the risk of failure. With that objective in mind, the idea is to adapt RSM in a functional context.

4.1.3 Organization of the chapter

We introduce in Section 4.2 an adaptation of RSM to functional data. Then, the methods we propose are studied numerically in Section 4.3. Finally, the Design of Experiments methods are applied to the CEA dataset in Section 4.4.

1. Commissariat à l'énergie atomique et aux énergies alternatives `http://www.cea.fr/`

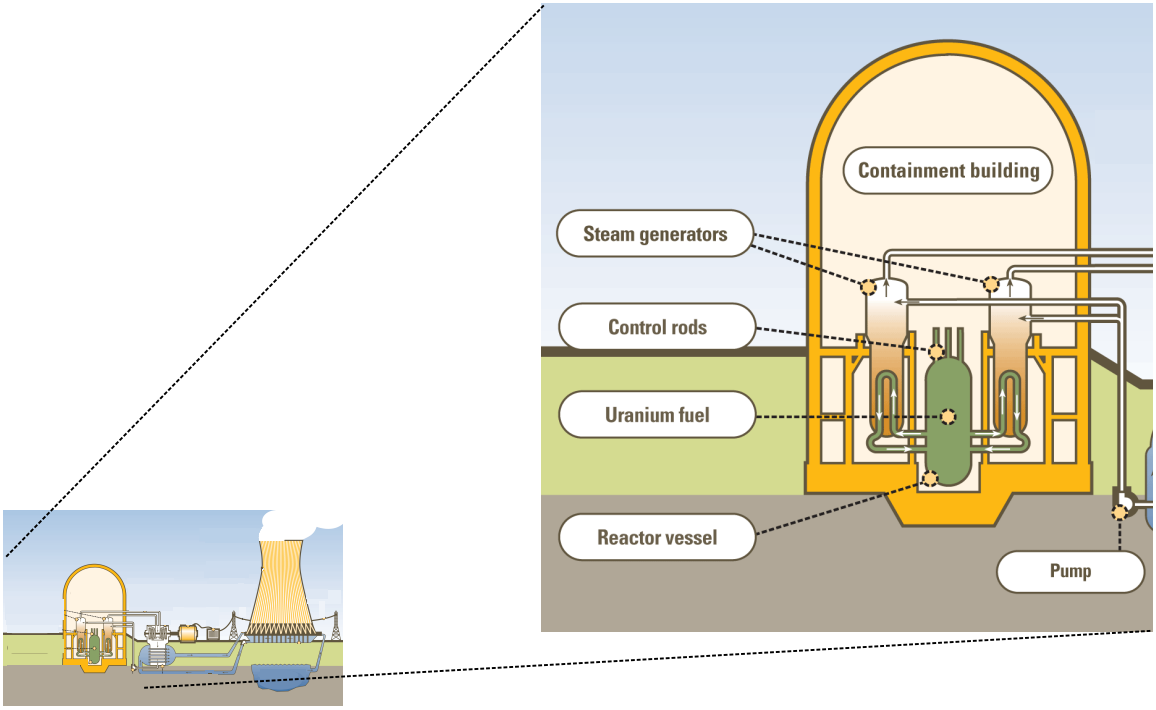


Figure 4.5: Diagram of a nuclear reactor. The primary circuit is in green.

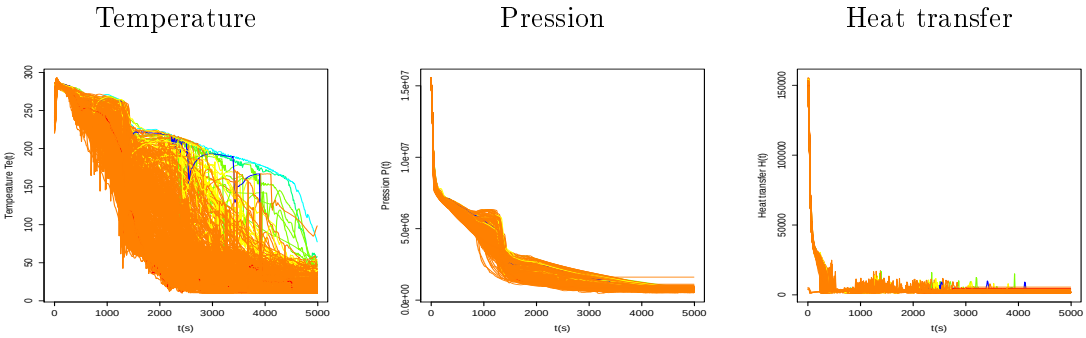


Figure 4.6: Evolution of temperature, pression and heat transfer. Red: $MF < 1$, orange: $1 \leq MF < 2$, yellow: $2 \leq MF < 3$, green: $3 \leq MF < 6$, light blue: $6 \leq MF < 7$ and dark blue $MF \geq 7$. Source: CEA.

4.2 Response Surface Methodology for functional data

We propose in this section an adaptation of Response Surface Methodology for functional data. The originality of the method is to provide functional designs of experiments.

4.2.1 Algorithm

Algorithm 1 is directly inspired by classical RSM procedures. The methods to generate the design and to calculate the least-squares estimators are detailed in sections 4.2.2 and 4.2.3 below.

Algorithm 1: Response Surface Algorithm

Initialization : Choose a starting point $x_0^{(0)}$;
for $k = 0, \ldots, k_{\max}$ **do**
  Generate a new set of experimental conditions : $\{x_1^{(k)}, ..., x_{n_k}^{(k)}\}$ around the starting point $x_0^{(k)}$;
  Obtain the corresponding responses : $\{Y_1^{(k)}, \ldots, Y_{n_k}^{(k)}\}$;
  Approximate the response surface using a linear model $Y_i^{(k)} := \alpha^{(k)} + \langle \beta^{(k)}, x_i^{(k)} \rangle + \varepsilon_i^{(k)}$, the least-squares estimator $(\widehat{\beta}^{(k)})$ indicates the direction of steepest-descent;
  Select a point $x_0^{(k+1)} = x_0^{(k)} - \alpha_k \widehat{\beta}^{(k)}$ $(\alpha_k > 0)$ in the direction of steepest descent;
end
We can stop the algorithm here and return $x_0^{(k_{\max})}$. *Or we can try to improve the optimization by fitting a second-order model as follows*;
Generate a new set of experimental conditions : $\left\{x_1^{(k_{\max})}, ..., x_{n_{k_{\max}}}^{(k_{\max})}\right\}$ around the point $x_0^{(k_{\max})}$;
Obtain the corresponding responses : $\left\{Y_1^{(k_{\max})}, \ldots, Y_{n_{k_{\max}}}^{(k_{\max})}\right\}$;
Approximate the response surface using a quadratic model $Y_i^{(k_{\max})} := \alpha_{k_{\max}} + \langle \beta^{(k_{\max})}, x_i^{(k_{\max})} \rangle + \frac{1}{2} \langle H x_i^{(k_{\max})}, x_i^{(k_{\max})} \rangle + \varepsilon_i^{(k_{\max})}$ and obtain least-squares estimators $(\widehat{\alpha}_{k_{\max}}, \widehat{\beta}^{(k_{\max})}, \widehat{H})$;
Result: $x^* := -\widehat{H}^{-1} \widehat{\beta}^{(k_{\max})}$.

4.2.2 Generation of a functional design of experiment

General principle

We propose a method of dimension reduction coupled with classical multivariate designs. The main idea is the following: suppose that we want to generate a design around $x_0 \in \mathbb{H}$, we choose an orthonormal basis $(\varphi_j)_{j\geq 1}$ of $\mathbb{H}$, a dimension d and a d-dimensional design $\{\mathbf{x}_i, i = 1, \ldots, n\} = \{(x_{i,1}, \ldots, x_{i,d}), i = 1, \ldots, n\}$ around $0 \in \mathbb{R}^d$ and we define a functional design $\{x_i, i = 1, \ldots, n\}$ verifying

$$x_i := x_0 + \sum_{j=1}^{d} x_{i,j}\varphi_j.$$

The advantage of such a method is its flexibility: all multivariate designs and all basis of $\mathbb{H}$ can be used. Then, by choosing an appropriate design and an appropriate basis, we can generate designs satisfying some constraints defined by the context.

Choice of basis

The choice of the basis has a significant influence on the quality of design. According to the context, it is possible to use a fixed basis such as Fourier basis, spline basis, wavelet basis, histogram basis...

However, if we have a training sample $\{(X_i, Y_i), i = 1, \ldots, n\}$, it may be relevant to use the information of this sample to find a suitable basis. The data-driven bases existing in the literature are.

- The PCA basis (see Section 1.1.1) which is the basis of $\mathbb{H}$ verifying

$$\frac{1}{n}\sum_{i=1}^{n} \|X_i - \widehat{\Pi}_d X_i\|^2 = \min_{\Pi_d} \left\{ \frac{1}{n}\sum_{i=1}^{n} \|X_i - \Pi_d X_i\|^2 \right\},$$

  where $\widehat{\Pi}_d$ is the orthogonal projector on $\text{span}\{\varphi_1, \ldots, \varphi_d\}$ and the minimum on the right-hand side is taken over all orthogonal projectors on d-dimensional subspaces of $\mathbb{H}$. More details on the computation of the PCA basis can be found in Section 2.4.2.
- The PLS basis (Wold, 1975; Preda and Saporta, 2005) which permits to take into account the interaction between X and Y. It is computed iteratively by the procedure described in Algorithm 2. For theoretical results on the PLS basis in a functional context see Delaigle and Hall (2012) and references therein.

4.2.3 Least-squares estimation and design properties

Remark that all the properties described in Section 4.1.1.3 for multivariate designs depend on the considered model (in particular its order p) and on the estimation method of the coefficients of the model (least-squares). In general functional contexts, the data lie in an infinite-dimensional space, and least-squares estimation methods can not be used without regularization (see Section 1.3.2, p.13 in Chapter 1). Then the properties of the design must also depend on the regularization method chosen to estimate the coefficients of the model.

Algorithm 2: practical implementation of PLS basis (Delaigle and Hall, 2012, Section A.2)

Data: Training sample $\{(X_i, Y_i), i = 1, \ldots, n\}$

Initialization :

$$X_i^{[0]} = X_i - \frac{1}{n}\sum_{i=1}^{n} X_i,\ Y_i^{[0]} = Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i$$

for $j = 1, \ldots, d$ **do**

Estimate φ_j by the empirical covariance of $X_i^{[j-1]}$ and $Y_i^{[j-1]}$:

$$\varphi_j = \sum_{i=1}^{n} Y_i^{[j-1]} X_i^{[j-1]} \Big/ \left\| \sum_{i=1}^{n} Y_i^{[j-1]} X_i^{[j-1]} \right\|$$

Fit the models $Y_i^{[j-1]} = \beta_j \langle X_i^{[j-1]}, \varphi_j \rangle + \varepsilon_i^{[j]}$ and $X_i^{[j-1]} = \delta_j \langle X_i^{[j-1]}, \varphi_j \rangle + W_i^{[j]}$ by least-squares that is

$$\widehat{\beta}_j := \sum_{i=1}^{n} Y_i^{[j-1]} \langle X_i^{[j-1]}, \varphi_j \rangle \Big/ \sum_{i=1}^{n} \langle X_i^{[j-1]}, \varphi_j \rangle^2$$

and

$$\widehat{\delta}_j := \sum_{i=1}^{n} \langle X_i^{[j-1]}, \varphi_j \rangle X_i^{[j-1]} \Big/ \sum_{i=1}^{n} \langle X_i^{[j-1]}, \varphi_j \rangle^2$$

Define $X_i^{[j]} := X_i^{[j-1]} - \langle X_i^{[j-1]}, \varphi_j \rangle \widehat{\delta}_j$ and $Y_i^{[j]} := Y_i^{[j-1]} - \widehat{\beta}_j \langle X_i^{[j-1]}, \varphi_j \rangle$ the residuals of the two fitted models;

end

However, the designs generated by the method proposed in Section 1.3.2 lie in a finite-dimensional subspace $\text{span}\{\varphi_1, \ldots, \varphi_d\}$ which is known by the user. This simplifies both estimation procedures and extension of notions of orthogonality, rotatability and optimality of design to our functional context. We focus on first-order and second-order designs but the same reasoning applies to more complex models.

First-order model

For our functional design, we suppose that $\{(x_i, Y_i), i = 1, \ldots, n\}$ verifies a first-order model

$$Y_i := \alpha + \langle \beta, x_i \rangle + \varepsilon_i, \text{ for } i = 1, \ldots, n,$$

with $\alpha \in \mathbb{R}$, $\beta \in \mathbb{H}$.

Now recall that, for all $i = 1, \ldots, n$, $x_i = x_0 + \sum_{j=1}^{d} x_{i,j}\varphi_j$, then the first-order model can be rewritten

$$Y_i := \alpha + \langle \beta, x_0 \rangle + \sum_{j=1}^{d} x_{i,j} \langle \beta, \varphi_j \rangle + \varepsilon_i, \text{ for } i = 1, \ldots, n.$$

This model is a first-order multivariate model and switching to notations of Section 4.1.1.3 it can be written

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,d} \\ 1 & x_{2,1} & & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,d} \end{pmatrix}$$

and coefficients $\boldsymbol{\beta} = (\alpha + \langle \beta, x_0 \rangle, \langle \beta, \varphi_1 \rangle, \ldots, \langle \beta, \varphi_d \rangle)'$. Then least-squares estimates of the model parameters can be obtained directly and it is easily seen that all first-order properties of the multivariate design $\{\mathbf{x}, i = 1, \ldots, d\}$ are also verified by the functional design.

Second-order model

Now we can see that a similar conclusion holds for the second-order model, which can be written here

$$Y_i := \alpha + \langle \beta, x_i \rangle + \frac{1}{2} \langle H x_i, x_i \rangle + \varepsilon_i, \text{ for } i = 1, \ldots, n, \tag{4.2}$$

where $\alpha \in \mathbb{R}$, $\beta \in \mathbb{H}$ and $H : \mathbb{H} \to \mathbb{H}$ is a linear self-adjoint operator. Then, by definition of $x_i = x_0 + \sum_{j=1}^{d} x_{i,j}\varphi_j$ we have

$$Y_i = \alpha + \langle \beta, x_0 \rangle + \frac{1}{2} \langle H x_0, x_0 \rangle + \sum_{j=1}^{d} x_{i,j} \left(\langle \beta, \varphi_j \rangle + \langle H x_0, \varphi_j \rangle \right) + \frac{1}{2} \sum_{j,k=1}^{d} x_{i,j} x_{i,k} \langle H\varphi_j, \varphi_k \rangle + \varepsilon_i.$$

This model is a second-order linear model for the data $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ and all second-order properties of $\{\mathbf{x}_i, i = 1, \ldots, n\}$ extend to the functional design $\{x_i, i = 1, \ldots, n\}$.

4.3 Numerical experimentation

In this section, we set $\mathbb{H} = \mathbb{L}^2([0,1])$ and $\mathcal{R} = \mathbb{H}$ (unconstrained minimization).

4.3.1 Functional designs

We use here the functions *cube*, *ccd* and *bbd* of the package *rsm* (Lenth, 2009) of *R* to generate respectively 2^d factorial designs, Central Composite Designs (CCD) and Box-Behnken Designs (BBD).

Functional designs with Fourier basis

In this section, we set $\varphi_1 \equiv 1$ and for all $j \geq 1$, for all $t \in [0,1]$,

$$\varphi_{2j}(t) = \sqrt{2}\cos(2\pi jt) \text{ and } \varphi_{2j+1}(t) = \sqrt{2}\sin(2\pi jt).$$

The curves of the generated designs are given in Figure 4.7.

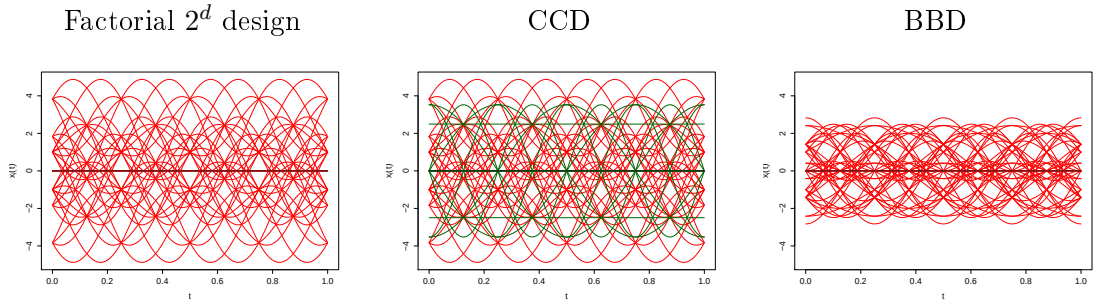


Figure 4.7: Functional designs with the Fourier basis ($d = 5$). Gray thick line: $x_0 \equiv 0$, red lines: points of the original 2^d or 3^d (for BBD) factorial design (see Section 4.1.1.2 p.131), green dotted lines: points added to the factorial design (for CCD). The colors of curves match the colors of points of figures 4.2, 4.3 and 4.4.

Functional design with data-driven bases

We simulate a sample $\{X_1, \ldots, X_n\}$ consisting of $n = 500$ realizations of the random variable

$$X(t) = \sum_{j=1}^{J} \sqrt{\lambda_j}\xi_j\psi_j(t),$$

with $J = 50$, $\lambda_j = e^{-j}$, $(\xi_j)_{j=1,\ldots,J}$ an i.i.d. sequence of standard normal random variables and $\psi_j(t) := \sqrt{2}\sin(\pi(j - 0.5)t)$.

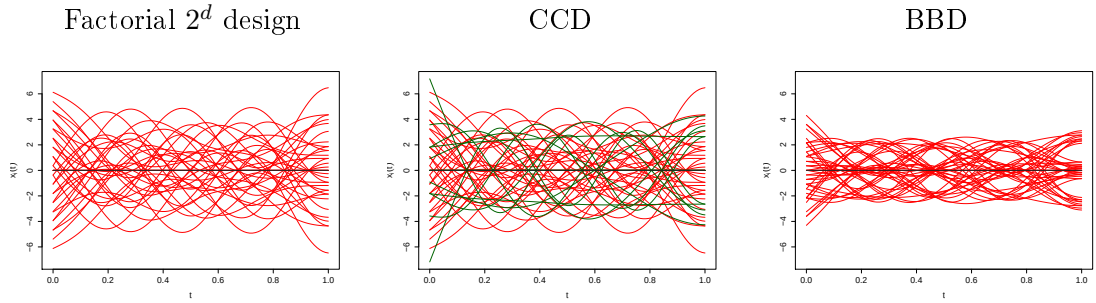


Figure 4.8: Functional designs with the PCA basis associated to $\{X_i, i = 1, \ldots, n\}$ $(d = 5)$. The legend is the same as the one of Figure 4.7.

The PCA basis only depends on $\{X_i, i = 1, \ldots, n\}$. In contrast, we will need to simulate the corresponding values of Y in order to calculate the PLS basis. In order to see the influence of the law of Y on the PLS basis we define two training samples $\left\{(X_i, Y_i^{(j)}), i = 1, \ldots, n\right\}$ for $j = 1, 2$ with

$$Y_i^{(j)} := m_j(X_i) + \varepsilon_i,$$

$m_j(x) := \|x - f_j\|^2$, where

$$\begin{aligned} f_1(t) &:= \cos(4\pi t) + 3\sin(\pi t) + 10, \\ f_2(t) &:= \cos(8.5\pi t)\ln(4t^2 + 10) \end{aligned}$$

and $\varepsilon_1, \ldots, \varepsilon_n$, i.i.d. $\sim \mathcal{N}(0, 0.01)$.

The curves of the design generated by the PLS basis (Figure 4.9) are much more irregular than those generated by the PCA basis (Figure 4.8). However, remark that the designs generated by the PLS basis (Figure 4.9) of the two samples show significant differences, which illustrates that the PLS basis effectively adapts to the law of Y.

4.3.2 Response surface algorithm

As an illustration, we run Algorithm 1 on the two examples given above. We use here the PLS basis calculated from the training sample $\{(X_i, Y_i^{(j)}), i = 1, \ldots, n\}$ with $j = 1$ or $j = 2$. The aim is to approach the minimum f_j of $m_j : x \in \mathbb{H} \to \|x - f_j\|^2$.

We use the training sample a second time to determine the starting point of the algorithm. We take

$$x_0^{(0)} := X_{i_{\min}} \text{ where } i_{\min} := \arg\min_{i=1,\ldots,n}\{Y_i\}.$$

The dimension is set to $d = 8$.

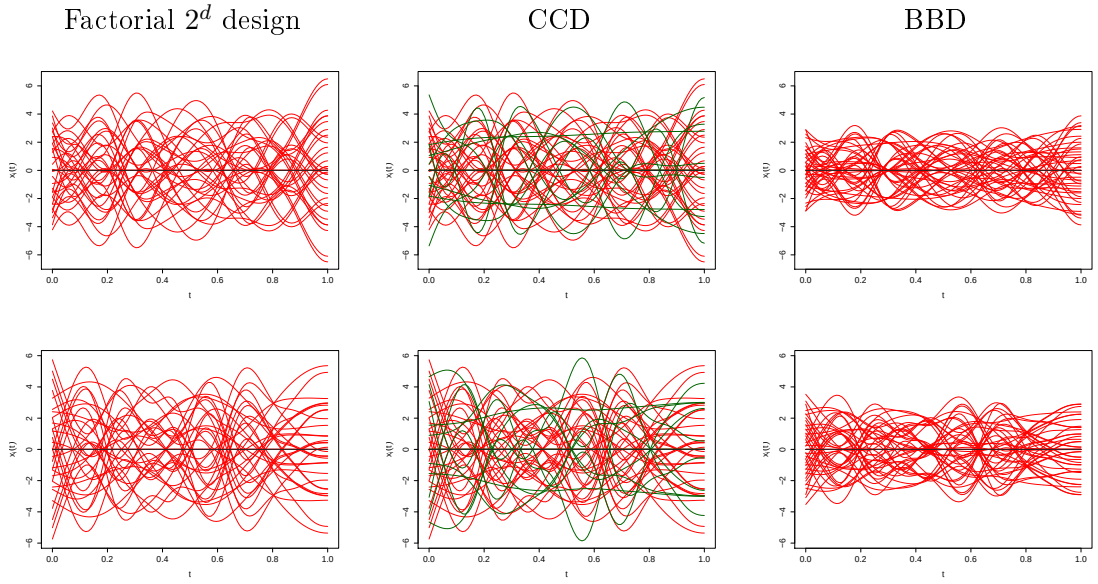


Figure 4.9: Functional designs with the PLS basis of the training sample $\{(X_i, Y_i^{(j)}, i = 1, \ldots, n\}$, $j = 1$ (first line) and $j = 2$ (second line), $d = 5$. The legend is the same as the one of Figure 4.7.

Approximation of $f_1 = \cos(4\pi t) + 3\sin(\pi t) + 10$.

Descent step. We generate a factorial 2^d design (Figure 4.9 – left) $(x_1^{(0)}, \ldots, x_{n_0}^{(0)})$ (here $n_0 = 2^d$) and we fit a first-order model

$$Y_i^{(0)} = \alpha^{(0)} + \sum_{j=1}^{d} \beta_j^{(0)} x_{i,j}^{(0)} + \varepsilon_i^{(0)},$$

to estimate the direction of steepest-descent. We realize two series of experiments on the direction of steepest descent. The first one (Figure 4.10– top left) allows us to suppose that the optimal value of α_0 is between 0.4 and 0.6 and the second one to fix $\alpha_0 = 0.50$. We set $m(x_0^{(1)}) := x_0^{(0)} - \alpha_0 \widehat{\beta}^{(0)}$.

The value of m at the starting point was $m(x_0^{(0)}) = 70.4 \pm 0.1$. At this step, we have $m(x_0^{(1)}) = 6.33 \times 10^{-3} \pm 10^{-5}$ and we have done only $2^d + 24 = 280$ experiments to reach this result.

We fit a first-order model once again with a 2^d factorial design and find that the norm of $\widehat{\beta}^{(1)}$ is very small ($\|\widehat{\beta}^{(1)}\| < 0.02$) compared to $\|\widehat{\beta}^{(0)}\| = 16.8 \pm 0.1$ which suggests that we are very close to a stationary point. We also note that the p-value of the Fisher's test $H_0 : \beta_1^{(1)} = \ldots = \beta_d^{(1)} = 0$ against $H_1 : \exists j \in \{1, \ldots, d\}, \beta_j^{(1)} \neq 0$ is very close to 1 which tends to confirm this assertion.

Final step. To improve the approximation, we fit a second-order model on the design points given by a Central Composite Design (Figure 4.9 – center). The matrix $\widehat{H}$ at this step is an estimation of the matrix of the restriction to the space $\text{span}\{\varphi_1, \ldots, \varphi_d\}$ of the Hessian operator of m at the point $x_0^{(1)}$. All the eigenvalues of $\widehat{H}$ are greater

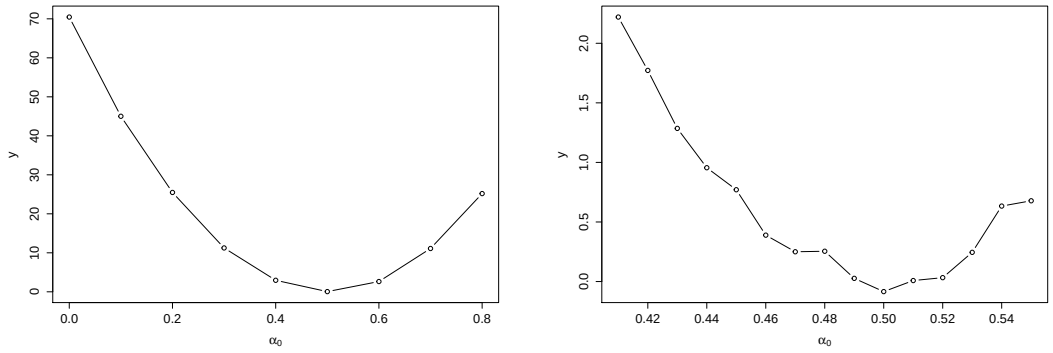


Figure 4.10: Results of experiments on the direction of steepest descent for the estimation of f_1. x-axis: α_0, y-axis: response $Y = m_1(x_0^{(0)} - \alpha_0\widehat{\beta}^{(0)}) + \varepsilon$.

than 1.96>0, this suggests that we are close to a minimum. We set $x_0^{(2)} := -\widehat{H}^{-1}\widehat{\beta}^{(1)}$ and we have $m(x_0^{(2)}) := 5.45 \times 10^{-3} \pm 10^{-5}$. The CCD with $d = 8$ counts 280 elements then we have realized 280 experiments for the descent step plus 280 for the final step, this rises to 557 the total number of experiments performed. Figure 4.11 represents the different results.

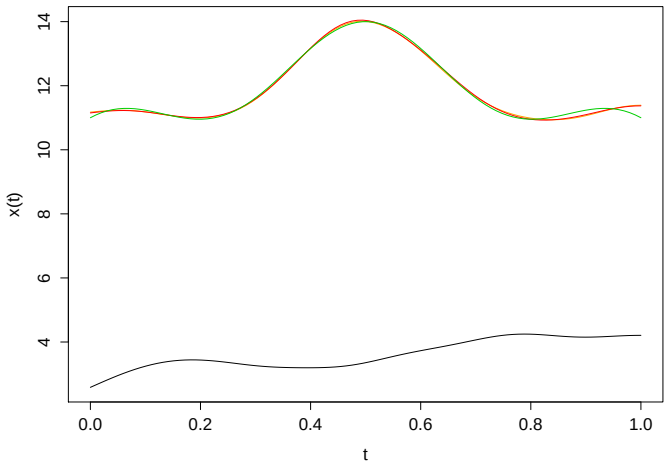


Figure 4.11: Result of optimization algorithm. Black curve: $x_0^{(0)}$, orange curve: $x_0^{(1)}$, red curve: $x_0^{(2)}$, green curve: f_1.

Approximation of $f_2(t) = \cos(8.5\pi t)\ln(4t^2 + 10)$

We have here $m(x_0^{(0)}) = 2.88 \pm 0.01$.

We follow the same steps as in the previous paragraph. Figure 4.12–left represents the evolution of the response along the direction of steepest descent. Here, since the response

is noisy, refining the result without doing a too large number of experiments seems to be difficult. Then, we fix $\alpha_0 = 0.5$ and $x_0^{(1)} = x_0^{(0)} - \alpha_0 \widehat{\beta}^{(0)}$. We have $m(x_0^{(1)}) = 1.99 \pm 0.01$. At this step, we have improved the response of about 31%. This is not as important as the improvement of the first step of estimation of f_1 but that is significant. This is probably due to the fact that the PLS basis is not optimal for generating good designs for the approximation of f_2. This fact highlights the importance of a good choice of basis and indicates that even a data-driven basis can not be optimal if the training sample is not well chosen.

This time, the p-value of the Fisher's test $H_0 : \beta_1^{(1)} = \ldots = \beta_d^{(1)} = 0$ against $H_1 : \exists j \in \{1, \ldots, d\}, \beta_j^{(1)} \neq 0$ is very small ($< 2 \times 10^{-4}$) which indicates that we are not close to a stationary point. Then, we try to improve the response doing a second descent step. However, the result of Figure 4.12–right seems to indicate that we will not improve significantly the response along this path.

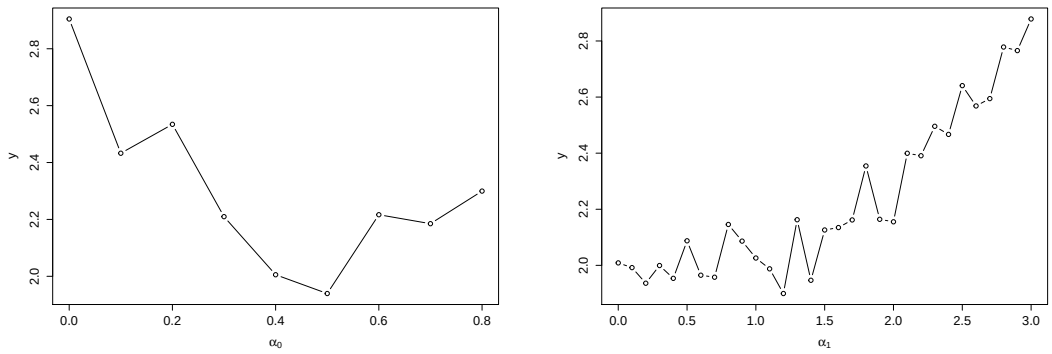


Figure 4.12: Results of experiments on the direction of steepest descent for the estimation of f_2. Left-hand side: first direction ($-\widehat{\beta}^{(0)}$), right-hand side: second direction ($-\widehat{\beta}^{(1)}$).

We fit a second-order model and set $x_0^{(2)} = -\widehat{H}^{-1}\widehat{\beta}^{(2)}$. We have $m(x_0^{(2)}) = 2.01 \pm 0.01$, compared to the result of the first step, we have not improved the response, probably because we are far from the stationary point.

4.3.3 Choice of basis

In this section, we compare the three bases proposed in Section 4.3.1 by a Monte-Carlo study.

We generate $n_s = 50$ training samples of size $n = 500$ and compare the results of the first descent step when the design is generated by the Fourier basis, the PCA basis and the PLS basis. The starting point is the same: $x_0^{(0)} = X_{i_{\min}}$ for $i_{\min} = \arg\min_{i=1,\ldots,n}\{Y_i\}$ (then for the Fourier basis the training sample is only used to set the starting point). The results are given in Figure 4.14. We see immediately that the PLS basis seems to be a better choice than the PCA one. However, the choice between the PLS basis and the Fourier basis is less clear and depends on the context.

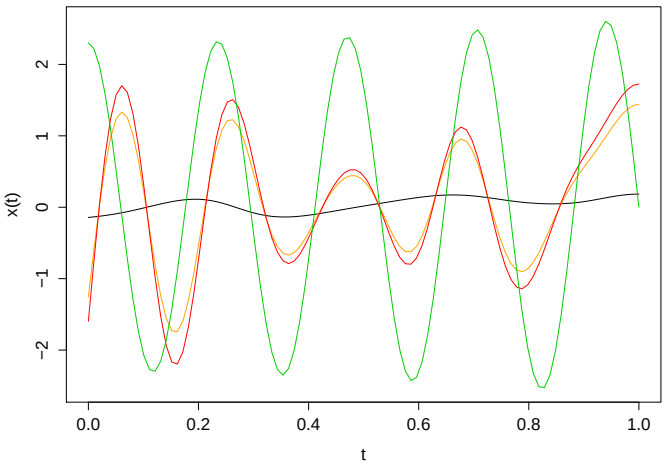


Figure 4.13: Result of optimization algorithm. Black curve: $x_0^{(0)}$, orange curve: $x_0^{(1)}$, red curve: $x_0^{(2)}$, green curve: f_2.

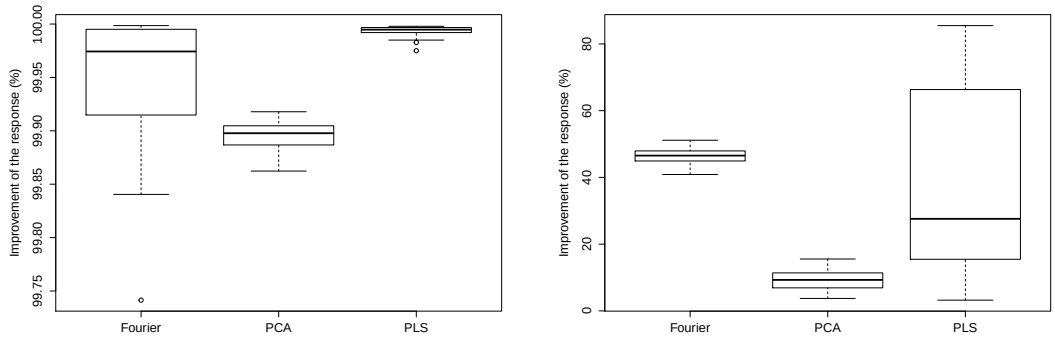


Figure 4.14: Monte-Carlo study of response improvement $\frac{m(x_0^{(0)})-m(x_0^{(1)})}{m(x_0^{(0)})}$ after the first descent step. Left-hand side: estimation of f_1, right-hand side: estimation of f_2.

4.3.4 Choice of dimension d

Figures 4.15 and 4.16 show that, except when the design is generated by the PCA basis for the approximation of f_1, the percentage of improvement increases when the dimension increases. Then, being aware that the number of experiments increases exponentially with the dimension, the user has to choose it as large as possible.

4.4 Data-driven experimental design for CEA dataset

The aim of this section is to generate a factorial design for the temperature, pressure and heat transfer transient given in Figure 4.6. Since the PLS basis has given good results in

simulations, we focus on this basis. We first generate the basis, whose three first components are plotted in Figure 4.17. This figure gives us some hints on the important features of the curves : for instance, we see that the behaviour of the temperature curves at the beginning and at the end of the simulation seems to have little influence on the margin factor since the three elements of the PLS basis are close to 0 in 0 and 5000. Conversely, the influence of the heat transfer on the margin factor is greater at the beginning of the simulation and seems to be negligible at the end.

Then, for each quantity considered (temperature, pressure, heat transfer) we define the starting point of the algorithm. We choose the element for which the margin factor is maximal (Figure 4.18).

For temperature and heat transfer, we generate a 2^5 factorial design around these initial curves. As some design points of the 2^5 factorial design generated around the initial heat transfer curve took negative values (which can not correspond to the physic since the heat transfer is always positive), we remove it and keep only the design points of the 2^5 factorial design which are always positive. The resulting design is a 2^{5-1} fractional factorial design.

If the final aim is to optimize margin factor with experiences involving variation of the three parameters (temperature transient, pressure transient and heat transfer transient), it is possible to generate a design by taking all the combination of the three designs represented in Figure 4.19. Since the number of combinations is large ($2^{5+5+4} = 16384$), it could be useful to choose $d = 3$ and have only $2^{3+3+2} = 256$ experiments to do in order to generate the sample.

At this step, we are not able to estimate a direction of steepest ascent since the computation code allowing to obtain the margin factors corresponding to the curves of the design generated is not public. Then, an immediate perspective of this work is to obtain the complete sample and run Algorithm 1 on these data.

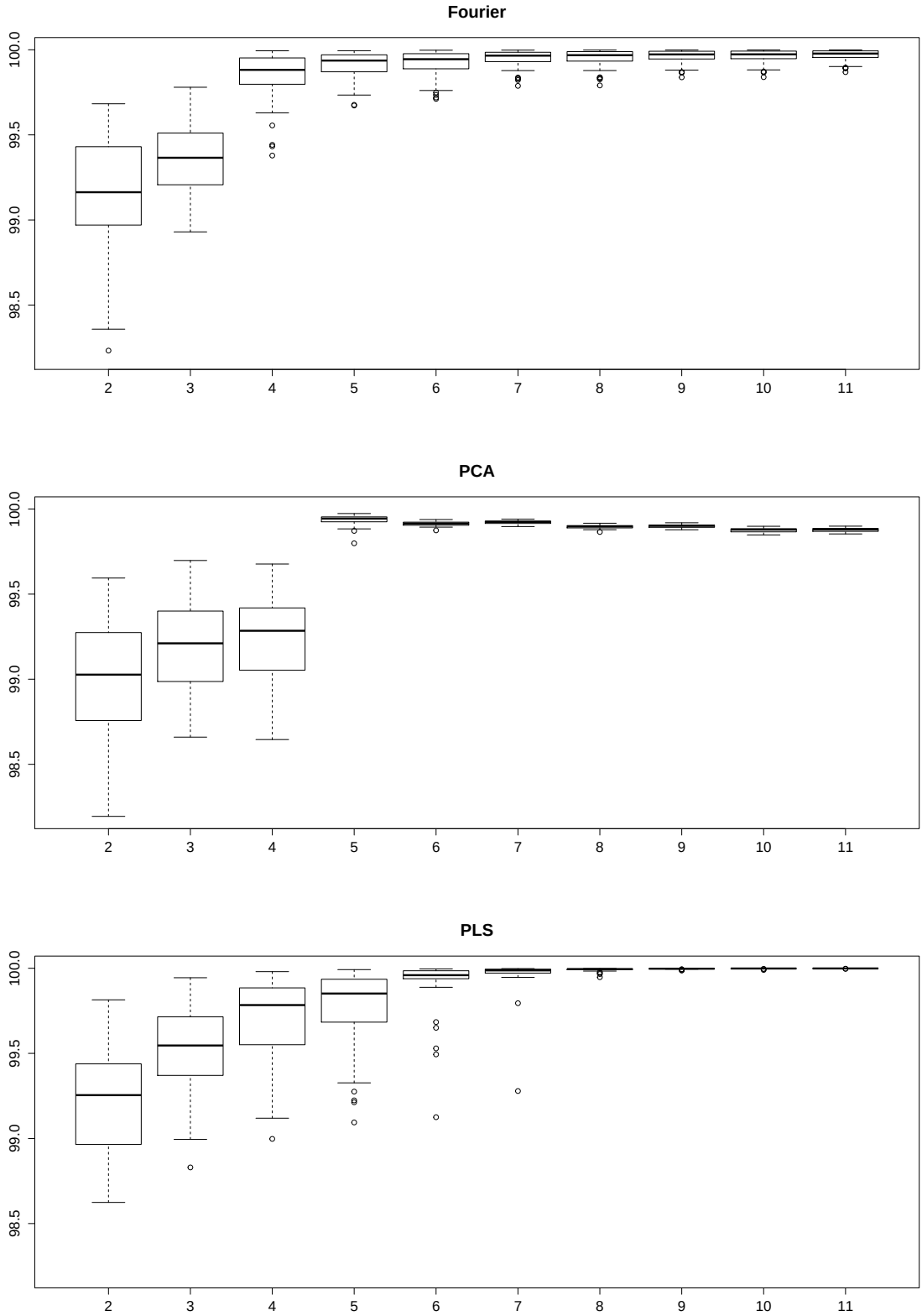


Figure 4.15: Monte-Carlo study of response improvement for the approximation of f_1 as a function of the dimension d.

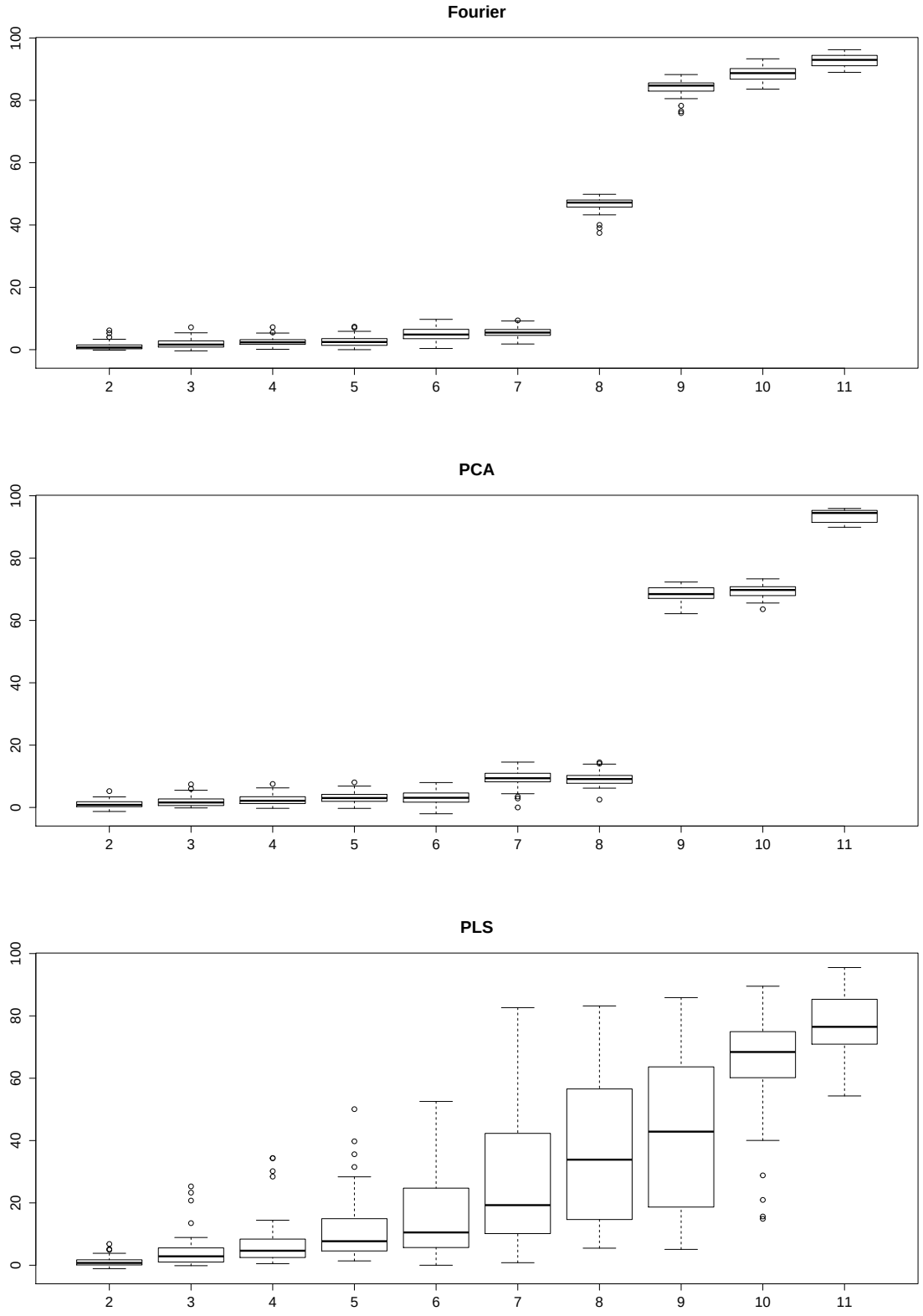


Figure 4.16: Monte-Carlo study of response improvement for the approximation of f_2 as a function of the dimension d.

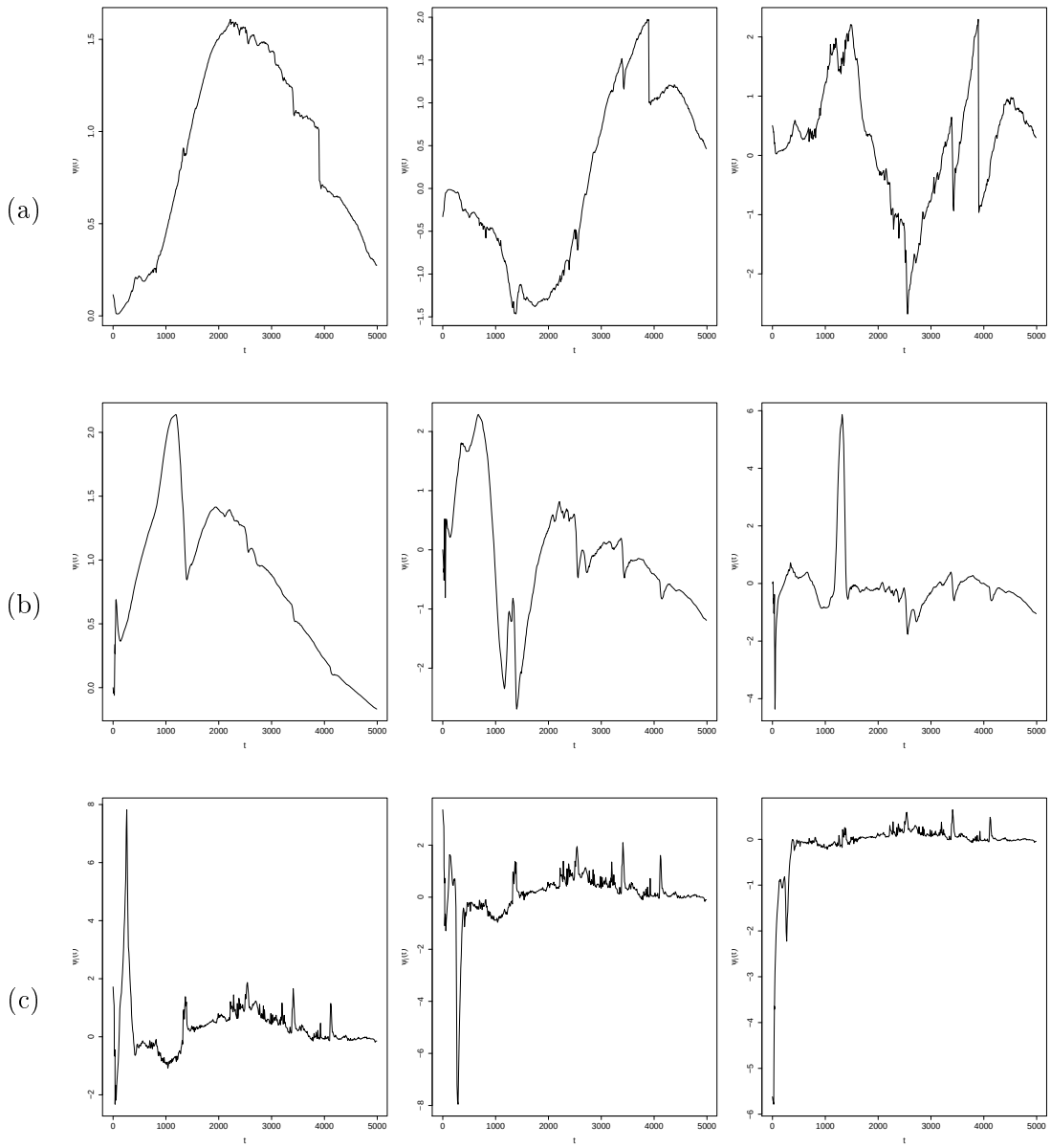


Figure 4.17: Three first functions of the PLS basis for the temperature/MF dataset (a), pressure/MF dataset (b) and heat transfer/MF (c)

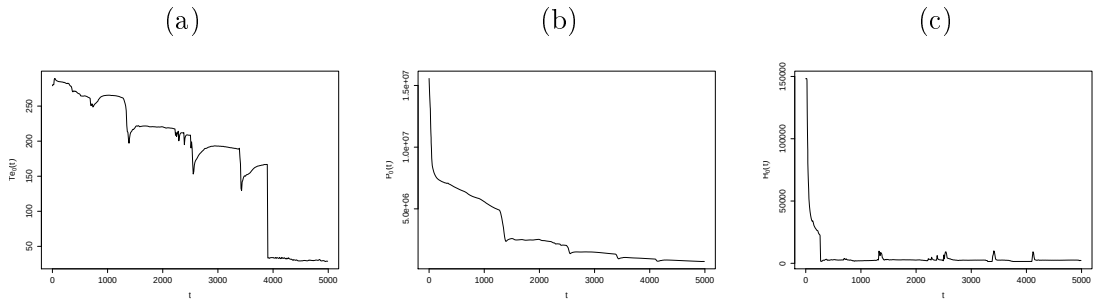


Figure 4.18: Initial temperature curve (a), pressure curve (b) and heat transfer curve (c).

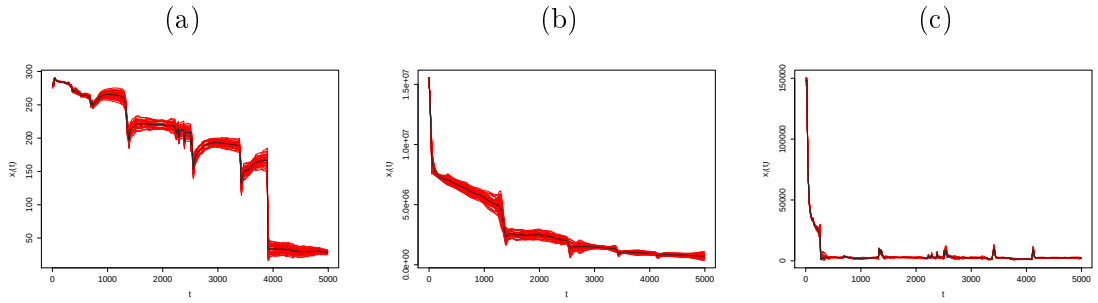


Figure 4.19: Factorial 2^5 design around the initial curves given by Figure 4.18 (a) temperature, (b) pressure. For heat transfer (c), the curve with negative values have been removed (fractional factorial 2^{5-1} design).

Conclusion et perspectives

Au cours de ce travail de thèse nous avons exploré des problématiques très diverses sur deux aspects : d'une part la nécessité de prendre en compte la spécificité des données fonctionnelles, liée à leur caractère infini-dimensionnel, ce qui nous a amené à nous questionner sur la réduction de la dimension et à explorer des outils spécifiques comme la théorie de la perturbation. D'autre part, dans l'objectif de proposer des procédures de sélection de modèle ou de fenêtre, nous avons adapté à ce contexte les idées de Birgé et Massart pour la sélection de modèle, et de Goldenshluger et Lepski pour la sélection de fenêtre, ce qui nous a permis d'obtenir des estimateurs adaptatifs dont le risque est contrôlé à taille d'échantillon finie.

Une problématique centrale dans ces travaux de thèse est la question de la réduction de la dimension. Dans chaque chapitre, cette question s'est posée et la réponse est de nature différente selon les contextes que nous avons considérés.

Dans le chapitre 2, nous avons premièrement constaté que l'estimation de la fonction de pente dans le modèle linéaire fonctionnel était un problème inverse mal posé. Dans ce cadre-là, projeter les données dans un espace de dimension finie bien choisi permet de régulariser ce problème inverse et d'obtenir des estimateurs satisfaisants. Une fois ce constat établi, se pose la question du choix de l'espace d'approximation qui recouvre en réalité deux problématiques : la question du choix de la base d'approximation et la question de la sélection de la dimension de l'espace.

De nombreux travaux ont traité du choix de la base d'approximation, que ce soit pour des problématiques liées à l'analyse des données fonctionnelles (voir par exemple Ramsay et Silverman 2005, Chapitre 3) ou de manière générale (DeVore et Lorentz, 1993). Lorsque la nature des données est connue, il est possible de définir des bases appropriées, comme la base de Fourier lorsque les données sont périodiques. Une autre solution est de définir une base "data-driven", comme la base de l'ACP, ce qui permet d'exploiter l'information de l'échantillon pour limiter la perte d'information lors de l'étape de projection. Une fois la base sélectionnée, la question de la sélection de la dimension se pose avec une importance particulière et, jusqu'à très récemment, peu de travaux traitaient de la sélection de la dimension en particulier d'un point de vue théorique. Dans la lignée des travaux de Comte et Johannes (2010), nous avons adapté la méthode de sélection de modèle avec contraste pénalisé au cadre fonctionnel. L'apport principal de nos travaux est de considérer non plus des bases fixes mais une base aléatoire, la base de l'ACP. Cette extension n'a pu se faire qu'avec l'aide de la théorie de la perturbation permettant de contrôler les projecteurs aléatoires de manière non-asymptotique. D'un point de vue pratique, la méthode de sélection de la dimension a montré des avantages par rapport aux méthodes usuelles de validation croisée, tant du point de vue du temps de calcul que du point de vue de la stabilité de l'estimateur final.

Bien que la question de la réduction de la dimension n'est pas la question centrale du chapitre 3, elle a été une des principales motivations à l'origine de ce travail. En effet,

rappelons la définition de l'estimateur de type Nadaraya-Watson de la fonction de répartition conditionnelle $F^x(y) = \mathbb{P}(Y \leq y|X = x)$ défini par Ferraty, Laksaci et Vieu (2006) et Ferraty, Laksaci, Tadj et al. (2010),

$$\widehat{F}_h^x(y) = \frac{\sum_{i=1}^n K_h(d(x, X_i))\mathbf{1}_{\{Y_i \leq y\}}}{\sum_{i=1}^n K_h(d(x, X_i))}, \tag{4.3}$$

avec d une pseudo-distance sur $\mathbb{H}$.

Il a été proposé, par exemple par Ferraty, Laksaci et Vieu 2006 ; Geenens 2011 de choisir des distances associées à des semi-normes de projection ($d(x, x') = \|\Pi_p(x - x')\|$ avec Π_p un projecteur sur un espace de dimension finie S_p) dans l'objectif de contourner le fléau de la dimension. L'hypothèse était que l'on pouvait obtenir, par ce biais, des estimateurs convergeant à une vitesse plus rapide que celle de l'estimateur (4.3) avec $d(x, x') = \|x - x'\|$. Remarquons que cela consiste à projeter les données dans l'espace S_p. Un des objectifs initiaux de ce travail était donc de s'intéresser à ces semi-normes, et plus précisément de sélectionner de manière adaptative les deux paramètres d'estimation : la dimension p et la fenêtre h. Au fur et à mesure de ce travail, il est apparu en réalité que, sous des hypothèses assez larges, lorsque la fenêtre h est bien choisie, la vitesse de convergence de l'estimateur (4.3) avec $d(x, x') = \|x - x'\|$ est optimale (cf. Tableau 3.1, p.84), ce qui implique qu'aucun estimateur ne peut converger à une vitesse plus rapide. Une analyse de la décomposition biais-variance de l'estimateur basé sur les semi-normes de projection montre que la variance est effectivement fortement réduite mais qu'en revanche le biais est augmenté à tel point que ces estimateurs ne peuvent atteindre la vitesse de convergence optimale que lorsque la base de l'espace d'approximation est bien choisie et que p tend vers l'infini à une certaine vitesse, ce qui pose des problèmes pratiques importants. Dans ce contexte, il semble donc que l'étape de projection des données dans un espace de dimension finie donne des estimateurs moins performants.

Dans le chapitre 4, la réduction de la dimension est vue comme une manière d'adapter au cadre fonctionnel des méthodes existant dans un cadre multivarié. Nous voyons, cependant, que la question du choix de la base d'approximation est, ici encore, une question centrale. En effet, rappelons que nous voulons optimiser une fonction différentiable $y = m(x)$ observée de manière bruitée (c'est-à-dire que, pour tout x, nous pouvons observer $y = m(x) + \varepsilon$ avec ε un bruit aléatoire). Nous partons d'un point x_0 et nous souhaitons estimer la direction de plus forte pente, donnée par le gradient de m en x_0 que nous noterons $m'(x_0)$. Pour cela, remarquant que, pour tout x, $m(x) = m(x_0) + \langle m'(x_0), x - x_0\rangle + o(\|x - x_0\|)$, l'idée est de choisir des points $(x_i)_{i=1,\ldots,n_0}$ proches de x, d'obtenir les résultats $(y_i)_{i=1,\ldots,n_0}$ correspondants et d'estimer $m'(x_0)$ par une régression linéaire de y_i par rapport à x_i. Nous avons la possibilité de choisir $(x_i)_{i=1,\ldots,n_0}$ et nous proposons le choix suivant : $x_i = x_0 + d_i$, avec d_i des points d'un espace de dimension finie S_d. Un tel choix implique que l'estimateur de $m'(x_0)$ obtenu est également dans l'espace S_d, et donc que l'espace S_d définit en quelque sorte un nombre restreint de directions d'optimisation. L'étude numérique du chapitre 4, nous indique que le choix de cet espace influence de manière importante le résultat de l'optimisation. Lorsque nous disposons d'un échantillon d'apprentissage, nous pouvons utiliser l'information de l'échantillon pour générer l'espace S_d. La régression PLS, qui permet de

générer une base tenant compte de la variabilité de Y par rapport à x, donne de bons résultats.

Une des perspectives de cette thèse serait de s'intéresser à des modèles de régression plus souples que le modèle linéaire fonctionnel mais pour lesquels nous pouvons espérer définir des procédures d'estimation convergeant à vitesse polynomiale. Les résultats de Mas (2012) montrent que ce n'est pas possible dans le cadre du modèle de régression non-paramétrique. Nous considérons donc des modèles intermédiaires en ajoutant une contrainte structurelle sur la fonction de régression.

Dans une première étape nous nous intéressons au modèle linéaire généralisé, que nous écrivons de la manière suivante :

$$Y = f(\alpha + \langle \beta, X \rangle) + \varepsilon, \tag{4.4}$$

avec ε un terme de bruit centré, de variance σ^2, indépendant de X, $f : \mathbb{R} \to \mathbb{R}$ une fonction supposée connue, α un nombre réel et β un élément de $\mathbb{H}$. Nous disposons d'un échantillon de $\{(X_i, Y_i),\ i = 1, \dots, n\}$ vérifiant l'équation (4.4) et l'objectif est d'estimer les paramètres (α, β).

Nous supposons que les conditions sont réunies pour que le modèle soit identifiable.

Une première motivation pour s'intéresser à ce modèle est d'ordre théorique, avec l'objectif de définir une procédure d'estimation adaptative pour le paramètre β, inexistante à ce jour à notre connaissance. D'autre part, la compréhension des problèmes liés à la non-linéarité du modèle serait un premier pas pour considérer des modèles plus généraux, comme le modèle single-index, où la fonction g est supposée inconnue.

Nous nous intéressons ici à des estimateurs par projection $(\widehat{\alpha}_m, \widehat{\beta}_m)$ définis par minimisation du contraste des moindres carrés

$$\gamma_n : (a, b) \in \mathbb{R} \times \mathbb{H} \mapsto \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(a - \langle b, X_i \rangle)\right)^2 \tag{4.5}$$

sur l'espace $\mathbb{R} \times S_m$, avec $S_m := \mathrm{Vect}\{\varphi_1, \dots, \varphi_{D_m}\}$ un espace d'approximation de dimension D_m défini à partir d'une base orthonormée fixe $(\varphi_j)_{j \geq 1}$ de $\mathbb{H}$.

Les premières simulations réalisées sur cet estimateur semblent donner de bons résultats représentés dans la figure 4.20, la base choisie est ici la base de Fourier. Sur les deux exemples présentés, un des estimateurs de la famille $\{\widehat{\beta}_m, m = 2, 5, 11, 17\}$ semble donner des résultats satisfaisants.

La question ici est de trouver un critère pour sélectionner la dimension D_m. L'idée est d'utiliser un critère de type moindres-carrés pénalisé, généralisant la procédure définie dans le chapitre 2. Pour cela, afin de déterminer la forme de la pénalité, nous souhaitons étudier la décomposition biais-variance de l'estimateur. Le problème principal est lié à la majoration de la quantité

$$\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (f(\widehat{\alpha} + \langle \widehat{\beta}_m, X_i \rangle) - f(\alpha + \langle \beta_m, X_i \rangle))\right], \tag{4.6}$$

indispensable pour obtenir l'ordre de la variance de l'estimateur (nous notons ici $\beta_m \in S_m$ la projection de β sur S_m).

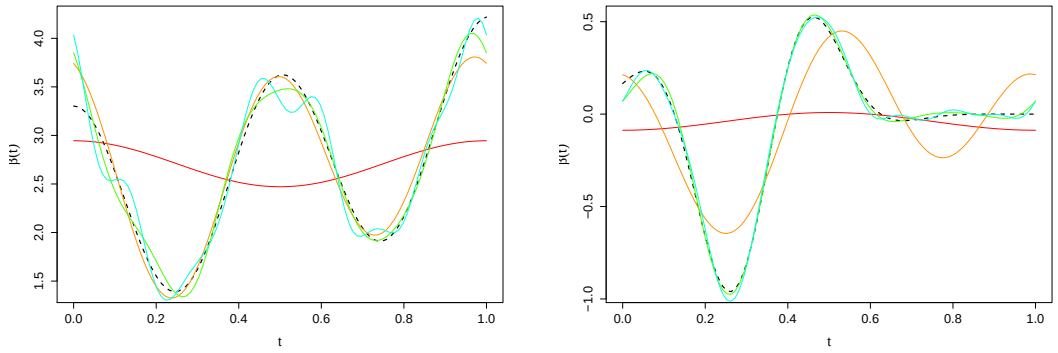


FIGURE 4.20 – Estimation de $\beta_1(t) := \ln(15t^2 + 10) + \cos(4\pi t)$ (à gauche) et de $\beta_2(t) := \exp(-(t-0.3)^2/0.05))\cos(4\pi t)$ (à droite). $n = 1000$, $\sigma^2 = 10$ (à gauche), $\sigma^2 = 0.001$ (à droite) pour $D_m = 2$ (courbes rouges), $D_m = 5$ (orange), $D_m = 11$ (vert) et $D_m = 17$ (bleu). Ici $g(t) = t^3$ et X est le mouvement brownien sur $[0, 1]$.

La non-linéarité pose ici un problème important. En effet, remarquons que les estimateurs $(\widehat{\alpha}_m, \widehat{\beta}_m)$ peuvent, de manière équivalente, être obtenus en minimisant la quantité

$$\widetilde{\gamma}_n : u \mapsto \frac{1}{n}\sum_{i=1}^{n}(Y_i - u(X_i))^2$$

sur l'ensemble

$$\mathcal{S}_m := \left\{ f : \mathbb{H} \to \mathbb{R}, \exists a \in \mathbb{R} \text{ and } (b_1, \ldots, b_{D_m})' \in \mathbb{R}^{D_m} \text{ tels que } f(x) = g\left(a + \sum_{j=1}^{D_m} b_j \langle x, \varphi_j \rangle\right) \right\}.$$

Sous cette écriture, nous voyons que la méthode d'estimation que nous proposons revient à minimiser le contraste des moindres carrés usuel en régression non-paramétrique mais sur un espace d'approximation qui, cette fois-ci, est non-linéaire. En effet, comme nous l'avons fait en régression linéaire fonctionnelle, nous pouvons ramener le problème de majoration de la quantité (4.6) au problème de majoration de

$$\mathbb{E}\left[\sup_{y=(y_1,\ldots,y_n)'\in B_m} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i y_i\right], \tag{4.7}$$

où B_m est l'intersection de la boule unité de $\mathbb{R}^n$ avec l'ensemble des éléments $(y_1, \ldots, y_n)'$ de $\mathbb{R}^n$ tels qu'il existe $f, f' \in \mathcal{S}_m$ vérifiant $y_i = f(X_i) - f'(X_i)$, pour tout i. La quantité (4.7) est liée à la topologie de l'ensemble B_m. Lorsque g est la fonction identité, l'ensemble B_m est la boule unité du sous-espace vectoriel S_m, de dimension D_m, et nous pouvons majorer cette quantité par $D_m\sigma^2/n$. La question qui se pose est donc d'adapter ce résultat en présence de non-linéarité.

Des espaces d'approximation non-linéaires se retrouvent également dans les modèles de

réseaux de neurones tels que définis par Barron (1994) sous la forme

$$\left\{ \sum_{j=1}^{D_m} g_j \ \phi(\langle a_j, x \rangle + b_j) + g_0 \right\},$$

avec $a_1, \ldots, a_{D_m} \in \mathbb{R}^k$, $g_0, g_1, \ldots, g_{D_m}, b_1, \ldots, b_{D_m} \in \mathbb{R}$ satisfaisant des contraintes que nous ne détaillerons pas ici. En introduisant une notion de dimension métrique, généralisant, entre autres, les travaux de Barron (1994) sur les modèles de réseaux de neurones, Birgé et Massart (1998) ont pu majorer des quantités de type (4.7) en présence de non-linéarité, ce qui pourrait constituer une piste pour débloquer le problème de majoration de la variance de notre estimateur.

Une deuxième motivation à l'étude des modèles linéaires généralisés est d'ordre pratique. Une discussion est en cours avec Lars Lau Rakêt et Stefan Sommer du département d'informatique de l'université de Copenhague dans l'objectif de définir une procédure pour diagnostiquer des troubles du déficit de l'attention avec hyperactivité (TDAH) à partir de données d'IRM fonctionnelle. Ces données, qui sont des images 3D du cerveau, prises toutes les 1.5 à 6 secondes peuvent être vues comme des données fonctionnelles dans le sens où ce sont des fonctions de $[0,T] \times [0,1]^d$, avec $d = 2$ ou 3, vers $\mathbb{R}$, que nous pouvons par exemple considérer comme étant des éléments de $\mathbb{L}^2([0,T] \times [0,1]^d, \mathbb{R})$.

Ce type de données nécessite un traitement approprié, il est par exemple nécessaire de supprimer les décalages éventuels liés à la manière dont les images sont enregistrées (recalage). D'autre part, les zones ayant un intérêt particulier sont des zones où un changement rapide en temps a lieu, il faut donc en tenir compte dans la méthode d'estimation des paramètres du modèle. Une collaboration avec un chercheur du laboratoire BRAINlab, dépendant du département de neurosciences et pharmacologie de l'université de Copenhague, est envisagée dans l'objectif de mieux comprendre la spécificité du problème considéré et d'adapter en conséquence à la fois le modèle et la méthode de prédiction. Nous envisageons pour le moment un modèle de type logit.

Outre les perspectives que nous venons d'évoquer, des axes de recherches faisant suite aux travaux de cette thèse sont envisageables.

- D'autres bases "data-driven", définies, par exemple, à partir de versions régularisées de l'ACP (Ramsay et Silverman, 2005, Chapitre 9), pourraient être considérées que ce soit dans le chapitre 2, de manière théorique, ou dans le chapitre 4.
  Les résultats du chapitre 2 pourraient être étendus à la base PLS (Preda et Saporta, 2005), qui a donné des résultats numériques satisfaisants dans le chapitre 4. L'intérêt pratique de cette base est de prendre en compte la variation de Y par rapport à X, ce qui n'est pas le cas de la base de l'ACP qui est calculée uniquement à partir des X_i. Toutefois, cet avantage pratique complique l'étude théorique des estimateurs définis sur la base PLS car cela interdit tout raisonnement basé sur un conditionnement par rapport à X, qui constituait une des clefs des résultats du chapitre 2.
  Une autre perspective du chapitre 2 serait de considérer d'autres méthodes d'estimation de la fonction de pente comme la régression *ridge* (Cardot, Ferraty et Sarda, 2003 ; Ramsay et Silverman, 2005) consistant à définir un estimateur minimisant le

critère

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle f, X_i\rangle)^2 + \rho\|f\|^2.$$

Pour cette méthode d'estimation, le paramètre ρ est un paramètre de lissage jouant le même rôle que la dimension D_m pour les estimateurs par projection. Une procédure de sélection basée sur une adaptation de la méthode de Goldenshluger et Lepski (2011) pourrait être envisagée dans ce cadre-là.

- Les résultats du chapitre 3 pourraient être étendus à d'autres problèmes d'estimation que celui de la fonction de répartition conditionnelle comme celui de la fonction de régression, de la densité conditionnelle ou de la fonction de hasard conditionnelle.
  La présence de censure sur Y ou de dépendance pourrait également être considérée.
- Une autre perspective du chapitre 3 serait d'explorer plus précisément la question de la semi-norme des estimateurs à noyaux. Nous avons établi que le remplacement de la norme dans le noyau par une semi-norme projective ne pouvait pas améliorer la vitesse de convergence de l'estimateur. Ce résultat est basé sur une hypothèse de type Hölder faisant intervenir la norme $\mathbb{L}^2$. Or, dans certaines situations, la norme $\mathbb{L}^2$ n'est pas la bonne norme à considérer, par exemple lorsque l'échantillon contient des courbes mesurées à des temps différents, deux courbes de l'échantillon peuvent avoir une distance $\mathbb{L}^2$ importante mais représenter le même processus. Partant de ce constat, il semblerait logique de considérer une semi-norme permettant de considérer deux courbes dont l'une est légèrement déformée par rapport à l'autre comme égales, une telle définition dépendrait bien sûr du contexte. Geenens (2011) propose un tel exemple de semi-norme dans l'objectif de faire de la reconnaissance automatique de signature.
  Ce type de modélisation pourrait nous permettre de contourner le fléau de la dimension. En effet, si l'on peut établir que chacune des courbes de l'échantillon est une déformation paramétrique aléatoire d'une certaine forme moyenne (Charlier, 2011) et que l'on peut tenir compte de cette déformation dans la définition de l'estimateur, il pourrait être possible d'obtenir des estimateurs à noyau convergeant à des vitesses plus rapides.

Annexe A

Théorie de la perturbation : application au contrôle des projecteurs aléatoires

L'objectif de la théorie de la perturbation est de comprendre le comportement des éléments propres d'un opérateur obtenu en "perturbant" (dans un sens à préciser) un autre opérateur. Cette théorie a été initialement développée par Rayleigh pour calculer les fréquences naturelles et les modes d'un système en vibration par comparaison avec ceux d'un système plus simple, et par Schrödinger pour des applications en mécanique quantique (voir l'introduction de Kato 1995 et les références qui y sont citées). Elle s'est ensuite révélée être un outil puissant en statistique pour contrôler les déviations des valeurs et vecteurs propres d'un opérateur ou d'une matrice de covariance empirique par rapport à ceux de sa version théorique (voir par exemple Koltchinskii et Giné 2000 ; Zwald et Blanchard 2005 ; Cardot, Mas et Sarda 2007 ; Crambes et Mas 2012). Nous présentons la théorie dans un premier temps, puis nous l'appliquons pour démontrer les résultats de contrôle non-asymptotique des composantes principales de l'échantillon sur lesquels sont basés les résultats du chapitre 2.

A.1 Mathematical tools

A.1.1 Operator theory

Let us first recall some notions of operator theory.

Definition 1. (Rudin 1973, Chapter 4 or Brezis 2005, Chapter VI)
Let $(\mathbb{H}, \|\cdot\|, \langle\cdot,\cdot\rangle)$ *be a Hilbert space.*

- *An operator* $T : \mathbb{H} \mapsto \mathbb{H}$ *is said to be* self-adjoint, *if for all* $x, y \in \mathbb{H}$,

$$\langle Tx, y\rangle = \langle x, Ty\rangle.$$

- *An operator* $T : \mathbb{H} \to \mathbb{H}$ *is said to be* compact *if the closure of the image by* T *of the open unit ball of* $\mathbb{H}$ *is compact.*
- *An operator* $T : \mathbb{H} \to \mathbb{H}$ *is said to be* Hilbert-Schmidt *if there exists an orthonormal basis* $(e_j)_{j\geq 1}$ *of* $\mathbb{H}$ *such that*

$$\sum_{j\geq 1} \|Te_j\|^2 < +\infty.$$

If an operator T is Hilbert-Schmidt, the quantity $\sum_{j\geq 1}\|Te_j\|^2$ is independent of the basis $(e_j)_{j\geq 1}$ and we can define a norm

$$\|T\|_{HS} := \sqrt{\sum_{j\geq 1}\|Te_j\|^2}.$$

For our purpose, the interest of the Hilbert-Schmidt norm is twofold. First this norm is associated to a scalar product

$$\langle T, T'\rangle_{HS} := \sum_{j\geq 1}\langle Te_j, T'e_j\rangle$$

and the space of all Hilbert-Schmidt operator equipped with this scalar product is a Hilbert space. Second, this norm may be easier to calculate than the more classical *operator norm* $\|\cdot\|_\infty$ defined by

$$\|T\|_\infty = \sup_{x\in\mathbb{H}}\frac{\|Tx\|}{\|x\|}$$

and the two norms are linked with the following relationship:

$$\|\cdot\|_\infty \leq \|\cdot\|_{HS}$$

(for more details see Bosq (2000, Chapter 1.5)). Recall that the space of all bounded linear operators equipped with the operator norm is a Banach space.

We admit here that all Hilbert-Schmidt operators are compact. We can find arguments and references for a proof of this result in Brezis (2005, p. 99).

We also recall the spectral theorem for self-adjoint compact operators.

Theorem 9. (Brezis 2005, Théorème VI.11, see also Halmos 1963) *Suppose that $\mathbb{H}$ is separable. Let T be a self-adjoint compact operator. Then $\mathbb{H}$ admits an orthonormal basis of eigenvectors of T.*

Examples: Both operators

$$\Gamma : f \in \mathbb{H} \mapsto \mathbb{E}[\langle f, X\rangle X]$$

and

$$\Gamma_n : f \in \mathbb{H} \mapsto \frac{1}{n}\sum_{i=1}^{n}\langle f, X_i\rangle X_i$$

are self-adjoint. Moreover, since Γ_n is a finite-rank operator, the image of the unit ball of Γ_n is a finite-dimensional bounded set. Then the closure of its image is a compact set and Γ_n is a compact operator.

It can also be seen that Γ is a Hilbert-Schmidt operator as soon as $\mathbb{E}\left[\|X\|^2\right] < +\infty$, which implies that Γ is a compact operator. Then, by Theorem 9, there exists a basis $(\psi_j)_{j\geq 1}$ of $\mathbb{H}$ (resp. $(\widehat{\psi}_j)_{j\geq 1}$) of eigenfunctions of Γ (resp. Γ_n). As a consequence, both

operators can be written in a spectral form

$$\Gamma : f \mapsto \sum_{j\geq 1} \lambda_j \langle f, \psi_j \rangle \text{ and } \Gamma_n : f \mapsto \sum_{j\geq 1} \widehat{\lambda}_j \langle f, \widehat{\psi}_j \rangle, \tag{A.1}$$

where $(\lambda_j)_{j\geq 1}$ and $(\widehat{\lambda}_j)_{j\geq 1}$ are the eigenvalues of Γ and Γ_n (sorted in decreasing order). Recall that the inverse of Γ is not a bounded operator and that Γ_n has no inverse (this can also be seen from Equation (A.1) since $\widehat{\lambda}_j = 0$ for $j > n$), we can define pseudo-inverses of both operators in the following way

$$\Gamma^\dagger : f \mapsto \sum_{j=1}^{K} \frac{\langle f, \psi_j \rangle}{\lambda_j} \psi_j \text{ and } \Gamma_n^\dagger : f \mapsto \sum_{j=1}^{K} \frac{\langle f, \hat{\psi}_j \rangle}{\hat{\lambda}_j} \hat{\psi}_j \mathbf{1}_{\{\hat{\lambda}_K > 0\}}, \tag{A.2}$$

with K a positive integer. The sums need to be finite to ensure that both $\Gamma^\dagger$ and $\Gamma_n^\dagger$ are continuous and well-defined for all $f \in \mathbb{H}$. These pseudo-inverse operators are also finite-rank then compact and self-adjoint for all $K > 0$.

Another consequence is that, for all $z \notin \{\lambda_j, j \geq 1\}$ (resp. $z \notin \{\widehat{\lambda}_j, j \geq 1\}$) the operators $(zI-\Gamma)^{-1}$, $(zI-\Gamma)^{-1/2}$, (resp. $(zI-\Gamma_n)^{-1}$, $(zI-\Gamma_n)^{-1/2}$) are well-defined and self-adjoint. For instance, since, for all $j \geq 1$, $(zI-\Gamma)^{-1/2}\psi_j = (\sqrt{z-\lambda_j})^{-1}\ \psi_j$ we have, for all $f \in \mathbb{H}$,

$$(zI-\Gamma)^{-1/2} f = \sum_{j\geq 1} \langle f, \psi_j \rangle (zI-\Gamma)^{-1/2} \psi_j = \sum_{j\geq 1} \frac{\langle f, \psi_j \rangle}{\sqrt{z-\lambda_j}} \psi_j.$$

A.1.2 Integration of Banach space valued functions

Let Q be a subset of $\mathbb{C}$ and E a complex Banach space equipped with a norm $\|\cdot\|_E$. We briefly define in this section the integration of continuous functions $f : Q \to E$ in a general framework. Then, in Section A.2, we will apply this definition to the particular case where E is the space $\mathcal{L}(\mathbb{H})$ of bounded operators from $\mathbb{H}$ to $\mathbb{H}$ equipped with the operator norm.

Let E^* be the dual space of E, that is the space of all bounded – for the operator norm – linear maps from E to $\mathbb{C}$.

Definition 2. (Rudin, 1973, Definition 3.26) *Suppose that E^* separate points that is to say that $\Lambda x = \Lambda y$ for all $\Lambda \in E^*$ implies $x = y$. Let $f : Q \to E$ be a continuous function such that, for all $\Lambda \in E^*$, Λf is integrable. If there exists $y \in E$ such that, for all $\Lambda \in E^*$,*

$$\Lambda y = \int_Q \Lambda f(z) dz,$$

then we define

$$y := \int_Q f(z) dz.$$

The existence of $\int_Q f(z)dz$ comes from Rudin (1973, Theorem 3.27) under the assumptions that Q is a compact Hausdorff space and the closure of the convex hull of $f(Q)$ is compact.

We can also define the notions of integrability and integral for E-valued functions as a limit of integrals of "simple" functions, generalizing the procedures used to define these

notions for $\mathbb{R}$-valued functions (*Bochner integral* see e.g. Dunford and Schwartz 1958). Both definitions coincides (see Dunford and Schwartz 1958, Section III.6.20).

Remark 7: The definition allows us to extend without much difficulty the classical theorems of complex analysis (Cauchy formula, residuals theorem,...) to E-valued functions.

A.1.3 Recalls on classical complex analysis

We recall that a *path* is a piecewise continuous differentiable function $\gamma : [a, b] \to \mathbb{C}$ where $[a, b] \in \mathbb{R}$ is an interval. We denote $\text{supp}(\gamma) := \{\gamma(t), t \in [a, b]\}$ and recall that the integral of a continuous function $f : \text{supp}(\gamma) \to \mathbb{C}$ with respect to γ is defined by

$$\int_\gamma f(\zeta)d\zeta = \int_a^b \gamma'(t)f(\gamma(t))dt.$$

We then recall the definition of the index of a complex number z with respect to a closed path γ.

Theorem 10. (Rudin, 1987, Theorem 10.10) *Let γ be a closed path. For all $z \notin \text{supp}(\gamma)$, define*

$$\text{Ind}_\gamma(z) := \frac{1}{2i\pi}\int_\gamma \frac{d\zeta}{\zeta - z}.$$

Then Ind_γ is an integer-valued function on $\mathbb{C}\backslash\text{supp}(\gamma)$ which is constant in each connected component of $\mathbb{C}\backslash\text{supp}(\gamma)$ and which is 0 on the unbounded connected component of $\mathbb{C}\backslash\text{supp}(\gamma)$.

The index $\text{Ind}_\gamma(z)$ essentially counts the "number of times γ winds around z". For simple paths, like the rectangular one (Figure A.1) or the circular one (Figure A.2), $\text{Ind}_\gamma(z)$ simply takes values in $\{0, 1\}$.

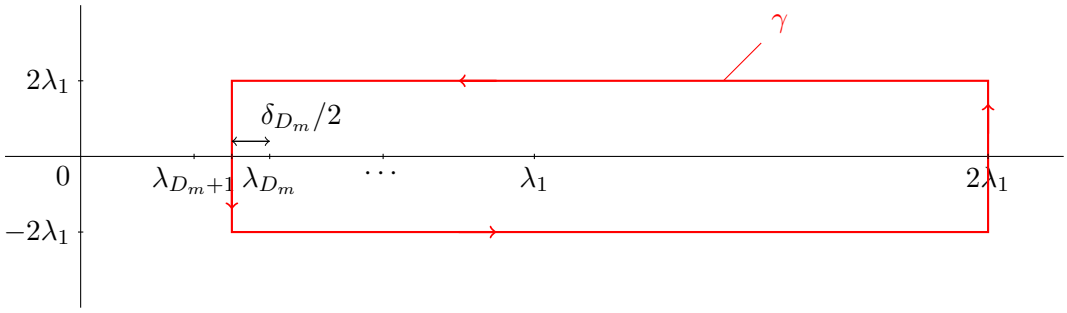


Figure A.1: Rectangular contour.

More precisely, for the rectangular path

$$\text{Ind}_\gamma(z) = \begin{cases} 1 & \text{if } \lambda_{D_m} - \delta_{D_m}/2 < \text{Re}\, z < 2\lambda_1 \text{ and } |\text{Im}\, z| < 2\lambda_1 \\ 0 & \text{otherwise} \end{cases},$$

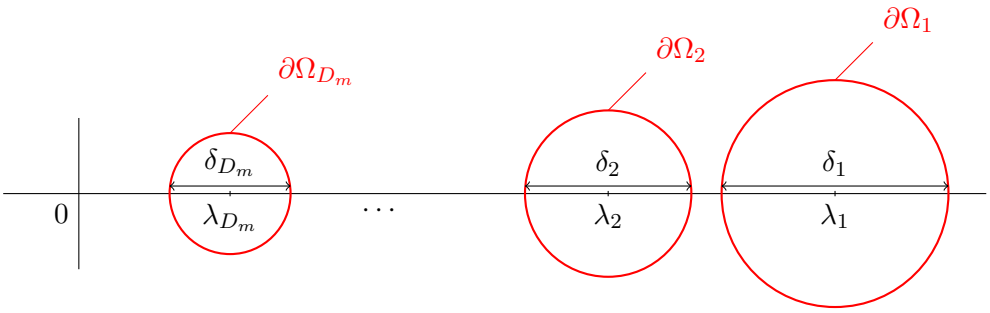


Figure A.2: Contour made of disjoint circles.

and for the circular path (see Rudin 1987, Theorem 10.11, p.204),

$$\mathrm{Ind}_\gamma(z) = \begin{cases} 1 & \text{if } |\lambda_j - z| < \delta_j/2 \text{ for a } j \in \{1, \dots, D_m\}, \\ 0 & \text{otherwise} \end{cases}.$$

Indeed, if γ is the circular path, the connected components of $\mathbb{C}\backslash\mathrm{supp}(\gamma)$ are the open disks $D(\lambda_j, \delta_j)$ of center λ_j and radius δ_j for $j = 1, \dots, D_m$ and $\mathbb{C}\backslash\bigcup_{j=1}^{D_m} \overline{D(\lambda_j, \delta_j)}$. If $z \in D(\lambda_j, \delta_j)$ for a $j = 1, \dots, D_m$ (in other words if $|z - \lambda_j| < \delta_j$), then, since Ind_γ is constant in $D(\lambda_j, \delta_j)$, we have:

$$\mathrm{Ind}_\gamma(z) = \mathrm{Ind}_\gamma(\lambda_j) = \frac{1}{2i\pi}\int_{\partial\Omega_j} \frac{d\zeta}{\zeta - \lambda_j} = \frac{1}{2i\pi}\int_0^1 \frac{2i\pi\delta_j e^{2i\pi t}}{\lambda_j + \delta_j e^{2i\pi t} - \lambda_j} dt = 1.$$

If $z \in \mathbb{C}\backslash\bigcup_{j=1}^{D_m} \overline{D(\lambda_j, \delta_j)}$, then it follows from Theorem 10 that $\mathrm{Ind}_\gamma(z) = 0$. The proof for the rectangular path is very similar.

We also recall the Cauchy Formula and the Residue Theorem.

Theorem 11. (Rudin, 1987, Theorem 10.15, p.207) *Suppose γ is a closed path in a convex open set Ω and f is holomorphic in Ω. If $z \in \Omega\backslash\mathrm{supp}(\gamma)$, then*

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

Theorem 12. (Rudin, 1987, Theorem 10.42, p.224) *Suppose that f is a meromorphic function in Ω. Let S be a set of points in Ω at which f has poles. If γ is a closed path in $\Omega\backslash S$ such that*

$$\mathrm{Ind}_\gamma(z) = 0 \text{ for all } z \notin \Omega,$$

then

$$\frac{1}{2i\pi}\int_\gamma f(\zeta)d\zeta = \sum_{\lambda\in S} \mathrm{Res}(f, \lambda)\mathrm{Ind}_\gamma(\lambda),$$

where $\mathrm{Res}(f, \lambda) = c_{-1}$ in the Laurent series expansion of f about the point λ,

$$f(z) = \sum_{j\geq -k} c_j(z - \lambda)^j.$$

A.2 Reformulation of random projectors with an integral

The aim of this section is to prove the following proposition, which is a key element in the proofs of lemmas 18, 19 and 20. This very important result allows us to write, on a set of high probability, the difference between theoretical and empirical projectors as an explicit function of the operators Γ and Γ_n. In the following, we need three assumptions (we keep here the notations of Chapter 2):

H2 There exists $b > 0$ such that, for all $\ell \in \mathbb{N}^*$,

$$\sup_{j\in\mathbb{N}} \mathbb{E}\left[\frac{< X, \psi_j >^{2\ell}}{\lambda_j^\ell}\right] \leq \ell! b^{\ell-1}.$$

H3 For all $j \neq k$, $< X, \psi_j >$ is independent of $< X, \psi_k >$.

H4 There exists a constant $\gamma > 0$ such that the sequence $\left(j\lambda_j \max\{\ln^{1+\gamma}(j), 1\}\right)_{j\geq 1}$ is decreasing.

Proposition 3. *Let γ be either the rectangular path given by Figure A.1 or the union (for $j = 1, ..., D_m$) of the circular paths $\partial\Omega_j$ of center λ_j and radius $\delta_j/2 := \min\{\lambda_j - \lambda_{j+1}, \lambda_{j-1} - \lambda_j\}/2$ represented in Figure A.2. There exists a set $\mathcal{A}_n$ such that*

$$\mathbb{P}\left(\mathcal{A}_n^{\complement}\right) \leq 2\exp(-c^* \ln^2 n),$$

where $c^* > 0$ depends on the sequence $(\lambda_j)_{j\geq 1}$ and

$$(\hat{\Pi}_m - \Pi_m)\mathbf{1}_{\mathcal{A}_n} = \frac{1}{2i\pi}\int_\gamma R^{1/2}(\zeta)\left[I - T(\zeta)\right]^{-1} T(\zeta) R^{1/2}(\zeta)\, d\zeta \mathbf{1}_{\mathcal{A}_n}, \tag{A.3}$$

$$(\hat{\pi}_j - \pi_j)\mathbf{1}_{\mathcal{A}_n} = \frac{1}{2i\pi}\int_{\partial\Omega_j} R^{1/2}(\zeta)\left[I - T(\zeta)\right]^{-1} T(\zeta) R^{1/2}(\zeta)\, d\zeta \mathbf{1}_{\mathcal{A}_n}, \tag{A.4}$$

$$(\Gamma^\dagger - \Gamma_n^\dagger)\mathbf{1}_{\mathcal{A}_n} = \frac{1}{2i\pi}\int_\gamma \frac{1}{\zeta} R^{1/2}(\zeta)\left[I - T(\zeta)\right]^{-1} T(\zeta) R^{1/2}(\zeta)\, d\zeta \mathbf{1}_{\mathcal{A}_n}, \tag{A.5}$$

where $\widehat{\Pi}_m$ is the orthogonal projector on Vect$\left\{\widehat{\psi}_1, \ldots, \widehat{\psi}_{D_m}\right\}$, Π_m is the orthogonal projector on Vect$\{\psi_1, \ldots, \psi_{D_m}\}$, $\widehat{\pi}_j$ is the orthogonal projector on Vect$\{\widehat{\psi}_j\}$, π_j is the orthogonal projector on Vect$\{\psi_j\}$, $\Gamma^\dagger$ and $\Gamma_n^\dagger$ are defined in Equation (A.2)*, p.160 and*

$$T(\zeta) := R^{1/2}(\zeta)(\Gamma_n - \Gamma)R^{1/2}(\zeta) \text{ and } R(\zeta) = (\zeta I - \Gamma)^{-1}.$$

Proof. First remark that, if E is the space of all the endomorphisms of $\mathbb{H}$, then, for all $j, k \geq 1$, the map $\Lambda_{j,k} := A \in E \mapsto \langle A\psi_j, \psi_k\rangle$ is in E^* and we can easily verify that, for all $A, A' \in E$,

$$\Lambda_{j,k}A = \Lambda_{j,k}A' \text{ for all } j, k \geq 1 \Rightarrow \Lambda A = \Lambda A' \text{ for all } \Lambda \in E^*.$$

Since Π_m is the orthogonal projector on span $\{\psi_1, \ldots, \psi_{D_m}\}$ and, for both contours, for

all j, $\mathrm{Ind}_\gamma(\lambda_j) = \mathbf{1}_{j\leq D_m}$, we have

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

and

$$\begin{aligned}\Lambda_{j,k}\Pi_m &= \langle\Pi_m\psi_j,\psi_k\rangle = \langle\frac{1}{2i\pi}\int_\gamma\frac{d\zeta}{\zeta-\lambda_j}\psi_j,\psi_k\rangle \\ &= \frac{1}{2i\pi}\int_\gamma\frac{d\zeta}{\zeta-\lambda_j}\mathbf{1}_{j=k} = \Lambda_{j,k}\frac{1}{2i\pi}\int_\gamma(\zeta I-\Gamma)^{-1}d\zeta,\end{aligned}$$

where $I : x \in \mathbb{H} \mapsto x$ is the identity operator. Then

$$\Pi_m = \frac{1}{2i\pi}\int_\gamma(\zeta I-\Gamma)^{-1}d\zeta, \tag{A.6}$$

by definition of $\int_\gamma(\zeta I-\Gamma)^{-1}d\zeta$. Now the aim is to write similarly the random projector $\hat{\Pi}_m$. This can be done if, for all $j \leq D_m$, $\hat{\lambda}_j$ is in the interior of γ.

We introduce the set $\mathcal{A}_n$ defined by

$$\mathcal{A}_n := \bigcap_{j=1}^{D_m}\left\{|\hat{\lambda}_j-\lambda_j| < \frac{\delta_j}{2}\right\}\bigcap\left\{\sup_{z\in\mathrm{supp}(\gamma)}\|T(z)\|_\infty < \frac{\mathbf{a}_k}{\sqrt{n}}\ln n\right\}, \tag{A.7}$$

where for all $k \geq 1$, $\mathbf{a}_k := \sum_{j\neq k}\frac{\lambda_j}{|\lambda_j-\lambda_k|} + \frac{\lambda_k}{\delta_k}$ and we recall that $T(\zeta) = R^{1/2}(\zeta)(\Gamma_n - \Gamma)R^{1/2}(\zeta)$.

We see that, on the set $\mathcal{A}_n$, following the same ideas as those of the proof of Equation (A.6), we also have that,

$$\widehat{\Pi}_m = \frac{1}{2i\pi}\int_\gamma(\zeta I-\Gamma_n)^{-1}d\zeta,$$

since $|\widehat{\lambda}_j-\lambda_j| < \delta_j/2$, for all $j \geq 1$ implies that $\mathrm{Ind}_\gamma(\widehat{\lambda}_j) = \mathbf{1}_{j\leq D_m}$. Remark that

$$\begin{aligned}(\zeta I-\Gamma)^{-1} - (\zeta I-\Gamma_n)^{-1} &= (\zeta I-\Gamma_n)^{-1}(\Gamma-\zeta I-(\Gamma_n-\zeta I))(\zeta I-\Gamma)^{-1} \\ &= (\zeta I-\Gamma_n)^{-1}R^{-1/2}(\zeta)T(\zeta)R^{1/2}(\zeta).\end{aligned} \tag{A.8}$$

Moreover $I - T(\zeta)$ is invertible on the set $\mathcal{A}_n$ for all $\zeta \in \mathrm{supp}(\gamma)$ (its minimal eigenvalue is greater than $1-\rho(T(\zeta)) \geq 1-\|T(\zeta)\|_\infty > 1/2$), and that

$$(I-T(\zeta))^{-1} = R^{-1/2}(\zeta)(\zeta I-\Gamma_n)^{-1}R^{-1/2}(\zeta),$$

or equivalently

$$(\zeta I-\Gamma_n)^{-1} = R^{1/2}(\zeta)(I-T(\zeta))^{-1}R^{1/2}(\zeta).$$

Then Equation (A.8) becomes

$$(\zeta I-\Gamma)^{-1} - (\zeta I-\Gamma_n)^{-1} = R^{1/2}(\zeta)(I-T(\zeta))^{-1}T(\zeta)R^{1/2}(\zeta),$$

and Equation (A.3) follows. Equation A.4 follows with a very similar proof.

For Equation (A.5), the proof is the same, once we have remarked that, for all $j \geq 1$,

$$\Gamma^{\dagger}\psi_j = \frac{1}{\lambda_j}\mathrm{Ind}_{\gamma}(\lambda_j)\psi_j = \frac{1}{2i\pi}\int_{\gamma}\frac{1}{\zeta(\zeta-\lambda_j)}d\zeta\psi_j,$$

where the last equality comes from Cauchy's Formula (Theorem 11) applied to the function $f : \zeta \mapsto \frac{1}{\zeta}$ with $z = \lambda_j$.

Now the proof is complete using the results of Lemma 16 hereafter. □

Lemma 16. *Suppose Assumption **H2** is fulfilled. There exists $c^* > 0$ depending only on the sequence $(\lambda_j)_{j\geq 1}$ and on b such that*

$$\mathbb{P}\left(\mathcal{A}_n^{\complement}\right) \leq 2\exp(-c^* \ln^2 n).$$

This result is a straightforward application of lemmas 13 (with $t = \mathbf{a}_j \ln n/\sqrt{n}$) and 14 of Mas and Ruymgaart (2014) (remarking that in our case $k \leq N_n \leq 20\sqrt{\frac{n}{\ln^3 n}}$). Both lemmas relies on a Bernstein type inequality for Hilbert-valued random variables (Lemma 23 p.180) and the minimax theorem for compact operator eigenvalues.

A.3 Upper-bound on the distance between empirical and theoretical projectors

We need a preliminary lemma which is a direct corollary of Hilgert, Mas, and Verzelen (2013, Lemma 10.1).

Lemma 17. *If Assumption **H4** is verified, then for all $k \in \mathbb{N}^*$:*

$$\mathbf{a}_k \leq C(\gamma)k\ln(k+1).$$

Lemma 18. *Let $r, R > 0$ and $\beta \in \mathcal{W}_r^R$. Suppose that assumptions **H2**, **H3** and **H4** are fulfilled. If $(\lambda_j)_{j\geq 1}$ decreases at polynomial rate (**P**) then*

$$\mathbb{E}[\|\widehat{\Pi}_m\beta - \Pi_m\beta\|_{\Gamma}^2] \quad \leq \quad C_1\frac{\ln^3(D_m)}{n}{D_m}^{\max\{(1-r)_+,2-a-r\}} + C_2\frac{\ln^9 n}{n^2}{D_m}^{(4+(a-r)_+-2a)_++2},$$

*and if $(\lambda_j)_{j\geq 1}$ decreases at exponential rate (**E**)*

$$\mathbb{E}[\|\widehat{\Pi}_m\beta - \Pi_m\beta\|_{\Gamma}^2] \leq C_2\frac{\ln^3(D_m)}{n}{D_m}^{(1-r)_+},$$

with $C_1 > 0$, $C_2 > 0$ and $C_3 > 0$ depending on r, R and on the sequence $(\lambda_j)_{j\geq 1}$ but are independent of m and n.

Proof. By Proposition 3, with γ the circular contour (Figure A.2)

$$\mathbb{E}[\|\widehat{\Pi}_m\beta - \Pi_m\beta\|_{\Gamma}^2\mathbf{1}_{\mathcal{A}_n^{\complement}}] \leq 2\|\beta\|_{\Gamma}^2\mathbb{P}(\mathcal{A}_n^{\complement}) \leq C\|\beta\|_{\Gamma}^2/n^2,$$

with C independent of β, n and m and

$$(\widehat{\Pi}_m\beta - \Pi_m\beta)\mathbf{1}_{\mathcal{A}_n} = \mathbf{1}_{\mathcal{A}_n}\frac{1}{2i\pi}\sum_{j=1}^{D_m}\int_{\partial\Omega_j} R^{1/2}(z)\left[I - T(z)\right]^{-1}T(z)R^{1/2}(z)\,\beta dz.$$

Remark that $(I-T(z))^{-1}T(z) = T(z) + (I-T(z))^{-1}T^2(z)$, then

$$\mathbf{1}_{\mathcal{A}_n}\left(\widehat{\Pi}_m - \Pi_m\right) = A_n + B_n,$$

with

$$A_n = \mathbf{1}_{\mathcal{A}_n}\frac{1}{2\pi i}\sum_{j=1}^{D_m}\int_{\partial\Omega_j}\left[(zI-\Gamma)^{-1}(\Gamma_n - \Gamma)(zI-\Gamma)^{-1}\right]dz,$$

$$B_n = \mathbf{1}_{\mathcal{A}_n}\frac{1}{2\pi i}\sum_{j=1}^{D_m}\int_{\partial\Omega_j}\left[(zI-\Gamma)^{-1/2}\left[I-T(z)\right]^{-1}T(z)^2(zI-\Gamma)^{-1/2}\right]dz.$$

We first deal with A_n. For all $j \geq 1$,

$$\langle A_n\beta, \psi_j\rangle = \sum_{k\geq 1}\beta_k\langle A_n\psi_k, \psi_j\rangle = \mathbf{1}_{\mathcal{A}_n}\frac{1}{2i\pi}\sum_{k\geq 1}\sum_{l=1}^{D_m}\int_{\Omega_l}\frac{\beta_k}{(\zeta-\lambda_j)(\zeta-\lambda_l)}\langle(\Gamma_n-\Gamma)\psi_k,\psi_j\rangle d\zeta,$$

where $\beta_j := \langle\beta, \psi_j\rangle$. By the Residue Theorem

$$\frac{1}{2i\pi}\int_{\Omega_l}\frac{d\zeta}{(\zeta-\lambda_j)(\zeta-\lambda_l)} = \begin{cases} \frac{1}{\lambda_j-\lambda_k} & \text{if } j = l \neq k \\ \frac{1}{\lambda_k-\lambda_j} & \text{if } k = l \neq j \\ 0 & \text{otherwise} \end{cases}.$$

Then

$$\langle A_n\beta,\psi_j\rangle = \mathbf{1}_{\mathcal{A}_n}\left(\sum_{k>D_m}\frac{\beta_k}{\lambda_j-\lambda_k}\mathbf{1}_{j\leq D_m}\langle(\Gamma_n-\Gamma)\psi_k,\psi_j\rangle + \sum_{k=1}^{D_m}\frac{\beta_k}{\lambda_k-\lambda_j}\mathbf{1}_{j>D_m}\langle(\Gamma_n-\Gamma)\psi_k,\psi_j\rangle\right).$$

Now remark that, by independence of $\langle X,\psi_j\rangle$ and $\langle X,\psi_k\rangle$ for $j\neq k$ and $\langle X_{i_1},\psi_j\rangle$ and $\langle X_{i_2},\psi_j\rangle$ if $i_1\neq i_2$, we have

$$\mathbb{E}\left[\langle(\Gamma_n-\Gamma)\psi_j,\psi_{k_1}\rangle\langle(\Gamma_n-\Gamma)\psi_j,\psi_{k_2}\rangle\right] = \frac{\lambda_j\lambda_k}{n}\mathbf{1}_{\{k_1=k_2=k\}}.$$

Then

$$\mathbb{E}[\langle A_n\beta,\psi_j\rangle^2] = \mathbf{1}_{\mathcal{A}_n}\frac{1}{n}\left(\lambda_j\mathbf{1}_{\{j\geq D_m\}}\sum_{k>D_m}\frac{\beta_k^2\lambda_k}{(\lambda_j-\lambda_k)^2} + \lambda_j\mathbf{1}_{\{j>D_m\}}\sum_{k=1}^{D_m}\frac{\beta_k^2\lambda_k}{(\lambda_k-\lambda_j)^2}\right),$$

$$\mathbb{E}\left[\|A_n\beta\|_\Gamma^2\right] = \mathbb{E}\left[\sum_{j\geq 1}\lambda_j\langle A_n\beta,\psi_j\rangle^2\right]$$
$$\leq \frac{1}{n}\sum_{j=1}^{D_m}\beta_j^2\sum_{k>D_m}\frac{\lambda_j\lambda_k^2}{(\lambda_j-\lambda_k)^2}+\frac{1}{n}\sum_{j=1}^{D_m}\lambda_j^2\sum_{k>D_m}\frac{\lambda_k\beta_k^2}{(\lambda_j-\lambda_k)^2}.$$

For the first term remark that

$$\sum_{k>D_m}\frac{\lambda_k^2}{(\lambda_j-\lambda_k)^2}=\left(\sum_{k>D_m}\frac{\lambda_k}{|\lambda_j-\lambda_k|}\right)^2-\sum_{k,l}^{D_m}\frac{\lambda_k\lambda_l}{|\lambda_j-\lambda_k||\lambda_j-\lambda_l|}\leq\left(\sum_{k>D_m}\frac{\lambda_k}{|\lambda_j-\lambda_k|}\right)^2\leq \mathbf{a}_j^2,$$

for the second since $j\leq D_m+1<k$, we have $\lambda_k\leq\lambda_{D_m+1}\leq\lambda_j$ and $\lambda_j-\lambda_k\geq\lambda_j-\lambda_{D_m+1}$, we obtain

$$\mathbb{E}\left[\|A_n\beta\|_\Gamma^2\right]\leq\frac{1}{n}\sum_{j=1}^{D_m}\lambda_j\beta_j^2\mathbf{a}_j^2+\frac{1}{n}\sum_{j=1}^{D_m}\frac{\lambda_j^2}{(\lambda_j-\lambda_{D_m+1})^2}\lambda_{D_m+1}{D_m}^{-r}\sum_{k>D_m}k^r\beta_k^2$$
$$\leq C\left(\frac{{D_m}^{(1-r)_+}\ln^3(D_m+1)}{n}+\frac{{D_m}^{2-r}\ln^2(D_m+1)}{n}\lambda_{D_m+1}\right), \qquad \text{(A.9)}$$

as soon as $(\lambda_j)_{j\geq 1}$ decreases exponentially or polynomially with $C>0$ depending only on R, r, a and Γ. The last line comes from the inequality $\lambda_j\leq Cj^{-1}$, Lemma 17 the fact that $\beta\in\mathcal{W}_r^R$ implies that $\beta_j\leq Cj^{-1-r}$ for a constant $C>0$ and that $\sum_{j=1}^{D_m}j^{-r}\ln^2(j+1)\leq C'\ln^3(D_m+1){D_m}^{(1-r)_+}$ for a constant $C'>0$.

We turn now to B_n

$$\|B_n\beta\mathbf{1}_{\mathcal{A}_n}\|_\Gamma^2$$
$$\leq\mathbf{1}_{\mathcal{A}_n}\frac{1}{4\pi^2}\sum_{k\geq 1}\lambda_k\left(\sum_{j=1}^{D_m}\int_{\partial\Omega_j}\langle R^{1/2}(\zeta)\left[I-T(\zeta)\right]^{-1}T(\zeta)^2R^{1/2}(\zeta)\beta,\psi_k\rangle d\zeta\right)^2. \qquad \text{(A.10)}$$

We have, since $R^{1/2}(\zeta)$ is self-adjoint,

$$\langle R^{1/2}(\zeta)\left[I-T(\zeta)\right]^{-1}T(\zeta)^2R^{1/2}(\zeta)\beta,\psi_k\rangle=\langle\left[I-T(\zeta)\right]^{-1}T(\zeta)^2R^{1/2}(\zeta)\beta,R^{1/2}(\zeta)\psi_k\rangle.$$

We denote by $\beta^\diamond:=\sum_{j\geq 1}j^{r/2}\beta_j\psi_j$; as β is in $\mathcal{W}_r^R$, the quantity $\|\beta^\diamond\|$ is finite and bounded by R. Moreover we denote by P_r the diagonal compact operator defined by $P_r\psi_j=j^{-r/2}\psi_j$, we remark that $\beta=P_r\beta^\diamond$. We have

$$R^{1/2}(\zeta)\psi_k=(\zeta I-\Gamma)^{-1/2}\psi_k=\frac{1}{\sqrt{\zeta-\lambda_k}}\psi_k.$$

Then,

$$|\langle R^{1/2}(\zeta)\left[I-T(\zeta)\right]^{-1}\widetilde{\Pi}(\zeta)^2 R^{1/2}(\zeta)\beta,\psi_k\rangle| \leq \frac{1}{\sqrt{|\zeta-\lambda_k|}}\left\|(I-T(\zeta))^{-1}\right\|_\infty \|T(\zeta)\|_\infty^2 \left\|R^{1/2}(\zeta)P_r\right\|_\infty \|\beta^\diamond\|.$$

Now on the set $\mathcal{A}_n$, by definition, we have for all $\zeta\in\Omega_j$:

$$\left\|(I-T(\zeta))^{-1}\right\|_\infty < 2 \text{ and } \|T(\zeta)\|_\infty < \frac{\mathbf{a}_j}{\sqrt{n}}\ln n.$$

Moreover, the eigenvalues of the operator $R^{1/2}(\zeta)P_r$ are $\{k^{-r/2}(\zeta-\lambda_k)^{-1/2}, k\geq 1\}$ then, for all, $\zeta\in\partial\Omega_j$:

$$\left\|R^{1/2}(\zeta)P_r\right\|_\infty = \sup_{k\geq 1}\{k^{-r/2}|\zeta-\lambda_k|^{-1/2}\}. \tag{A.11}$$

Now remark that, for all $k>j$, for all $\zeta\in\partial\Omega_j$,

$$|\zeta-\lambda_k| \geq \lambda_j - \delta_j/2 - \lambda_k.$$

and

$$k^{-r/2}|\zeta-\lambda_k|^{-1/2} \leq k^{-r/2}(\lambda_j-\delta_j/2-\lambda_k)^{-1/2} \leq j^{-r/2}\delta_j^{-1/2},$$

since the sequence $\left(k^{-r/2}(\lambda_j-\delta_j/2-\lambda_k)^{-1/2}\right)_{k>j}$ is decreasing and $\lambda_j-\lambda_{j+1}\geq\delta_j$. For $k=j$, $|\zeta-\lambda_k|=\delta_j/2$ ($\partial\Omega_j$ is the circle of center λ_j and radius $\delta_j/2$), hence $k^{-r/2}|\zeta-\lambda_k|^{-1/2} = j^{-r/2}\delta_j^{-1/2}$. Then, from Equation (A.11), we deduce that

$$\left\|R^{1/2}(\zeta)P_r\right\|_\infty \leq \sup_{1\leq k\leq j}\{k^{-r/2}|\zeta-\lambda_k|^{-1/2}\}.$$

Now, for $k=1,\ldots,j$, we have

$$|\zeta-\lambda_k| \geq \lambda_k-\lambda_j-\delta_j/2 \geq (\lambda_k-\lambda_j)/2,$$

since $\lambda_k-\lambda_j\geq\delta_j$ (by definition of δ_j) implies that $\frac{\lambda_j-\lambda_j-\delta_j/2}{\lambda_k-\lambda_j} = 1-\delta_j/(2(\lambda_k-\lambda_j)) \geq 1/2$. Then we have

$$\begin{aligned}\left\|R^{1/2}(\zeta)P_r\right\|_\infty &\leq \max\left\{\sup_{1\leq k\leq j-1}\left\{\lambda_k^{-1/2}k^{-r/2}\left(\frac{\lambda_k}{\lambda_k-\lambda_j}\right)^{1/2}\right\}, \lambda_j^{-1/2}j^{-r/2}\lambda_j^{1/2}\delta_j^{-1/2}\right\}\\ &\leq \mathbf{a}_j^{1/2}\sup_{1\leq k\leq j}\left\{\lambda_k^{-1/2}k^{-r/2}\right\},\end{aligned}$$

where we recall that $\mathbf{a}_j = \sum_{k\neq j}\frac{\lambda_k}{|\lambda_j-\lambda_k|}+\frac{\lambda_j}{\delta_j}$.

Now, in the polynomial case **(P)**,

$$\left\|R^{1/2}(\zeta)P_r\right\|_\infty \leq c_P^{-1/2}\mathbf{a}_j^{1/2}\sup_{1\leq k\leq j}\left\{k^{(a-r)/2}\right\} \leq c_P^{-1/2}\mathbf{a}_j^{1/2}j^{(a-r)_+/2}. \tag{A.12}$$

Then Equation (A.10) becomes:

$$\begin{aligned}\|B_n\beta\mathbf{1}_{\mathcal{A}_n}\|_\Gamma^2 &\leq \frac{R^2}{\pi^2}\sum_{k\geq 1}\lambda_k\left(\sum_{j=1}^{D_m}\frac{\mathbf{a}_j^{3/2}}{n}\ln^2 n\; j^{(a-r)_+/2}\int_{\partial\Omega_j}\frac{dz}{\sqrt{|z-\lambda_k|}}\right)^2\\ &\leq \frac{R^2}{\pi^2}\sum_{k\geq 1}\lambda_k\left(\sum_{j=1}^{D_m}\frac{\mathbf{a}_j^{3/2}}{n}\ln^2 n\; j^{(a-r)_+/2}\frac{\pi\delta_j}{\sqrt{D_{j,k}}}\right)^2,\end{aligned}$$

where $D_{j,k} = |\lambda_j - \lambda_k|$ if $j \neq k$ and $Dj,j = \delta_j$ (remark that, for all $z \in \partial\Omega_j$, $|z - \lambda_k| > D_{j,k}/2$). By Cauchy Schwarz's Inequality and by Lemma 17 which states that $\mathbf{a}_j \leq Cj\ \ln(j+1)$ (with $C > 0$ depending only on Γ)

$$\|B_n\beta\mathbf{1}_{\mathcal{A}_n}\|_\Gamma^2 \leq 2C^3C_P^2R^2\frac{\ln^7(n)}{n^2}\sum_{k\geq 1}\lambda_k\sum_{j=1}^{D_m}j^{3+(a-r)_+-2a}\sum_{j=1}^{D_m}D_{j,k}^{-1}.$$

Now $\sum_{k\geq 1}\lambda_k D_{j,k}^{-1} = \mathbf{a}_j \leq Cj\ \ln(j+1)$, then

$$\|B_n\beta\mathbf{1}_{\mathcal{A}_n}\|_\Gamma^2 \leq 2C^3C_P^2R^2\frac{\ln^8(n)}{n^2}\sum_{j=1}^{D_m}j^{3+(a-r)_+-2a}\sum_{j=1}^{D_m}j.$$

Remarking that, for all $\alpha \in \mathbb{R}$, $\sum_{k=1}^{D_m}k^\alpha \leq c_\alpha\ln(j+1)\ j^{(\alpha+1)_+}$, we get

$$\|B_n\beta\mathbf{1}_{\mathcal{A}_n}\|_\Gamma^2 \leq C'\frac{\ln^9 n}{n^2}D_m^{(4+(a-r)_+-2a)_++2},$$

with $C' > 0$ depends only on R, r, a, Γ, c_P and C_P.

Gathering with (A.9) we obtain the expected result.

In the exponential case **(E)** we have, by Lemma 17,

$$\left\|R^{1/2}(\zeta)P_r\right\|_\infty \leq c_E^{-1/2}\mathbf{a}_j^{1/2}\sup_{1\leq k\leq j}\left\{e^{k^a/2}k^{-r/2}\right\} \leq Cj^{1/2}\ln^{1/2}(j+1)\ e^{j^a/2}j^{-r/2} \qquad \text{(A.13)}$$

and following the same scheme as for the polynomial case (**P**), we get

$$\begin{aligned}\|B_n\beta\mathbf{1}_{\mathcal{A}_n}\|_\Gamma^2 &\leq C^3\frac{R^2}{\pi^2}\sum_{k\geq 1}\lambda_k\left(\sum_{j=1}^{D_m}\frac{j^{3/2}\ln^{3/2}(j+1)}{n}\ln^2(n)\ e^{j^a/2}j^{-r/2}\frac{\pi\delta_j}{\sqrt{D_{j,k}}}\right)^2\\ &\leq C^3R^2C_E^2\frac{\ln^7(n)}{n^2}\sum_{k\geq 1}\lambda_k\sum_{j=1}^{D_m}j^{6-r}e^{-j^a}\sum_{j=1}^{D_m}j^{-3}D_{j,k}^{-1}\\ &\leq C\frac{\ln^7 n}{n^2}\sum_{j=1}^{D_m}j^{-2}\ln(j+1) \leq \frac{C''}{n},\end{aligned}$$

since both series $\sum_{j\geq 1}j^{6-r}e^{-j^a}$ and $\sum_{j\geq 1}j^{-2}\ln j$ are convergent (here C', C'' are two positive real numbers depending only on R, r, a, Γ, c_E and C_E). □

A.4 Empirical and theoretical bias terms

Lemma 19. *Suppose that assumptions **H2**, **H3** and **H4** are fulfilled and that $\beta \in \mathcal{W}_r^R$ with $r, R > 0$ such that, in the polynomial case **(P)**, either $a + r/2 > 2$. Then for all $m = 1, ..., N_n$:*

$$\mathbb{E}[\|\beta - \widehat{\Pi}_m \beta\|_{\Gamma_n}^2] \leq 4\mathbb{E}[\|\beta - \widehat{\Pi}_m \beta\|_{\Gamma}^2] + C\frac{D_m}{n}, \tag{A.14}$$

where $C > 0$ is independent of β and m.

Proof. First imagine that the random projectors in the equation above are replaced by non random one. It is elementary to see that

$$\mathbb{E}\left[\|\beta - \Pi_m \beta\|_{\Gamma_n}^2\right] = \|\beta - \Pi_m \beta\|_{\Gamma}^2$$

and that consequently to get (A.14) it is enough to show that both $\mathbb{E}\left[\left\|\left(\Pi_m - \widehat{\Pi}_m\right)\beta\right\|_{\Gamma}^2\right]$ and $\mathbb{E}\left[\left\|\left(\Pi_m - \widehat{\Pi}_m\right)\beta\right\|_{\Gamma_n}^2\right]$ are bounded, up to a multiplicative constant, by D_m/n. The first bound was proved asymptotically in Cardot, Mas, and Sarda (2007) and non-asymptotically in Crambes and Mas (2012, Proposition 20) in a slightly more general framework. Specifically these authors get

$$\mathbb{E}\left[\left\|\left(\Pi_m - \widehat{\Pi}_m\right)\beta\right\|_{\Gamma}^2\right] \leq A\frac{{D_m}^2 \lambda_{D_m}}{n},$$

where A does not depend on m and n. The only point left is to prove the same sort of bound for

$$\mathbb{E}\left[\left\|\left(\Pi_m - \widehat{\Pi}_m\right)\beta\right\|_{\Gamma_n}^2\right].$$

In a first step remark that

$$\left\|\left(\Pi_m - \widehat{\Pi}_m\right)\beta\right\|_{\Gamma_n}^2 = \left\langle\left(\Gamma_n - \Gamma\right)\left(\Pi_m - \widehat{\Pi}_m\right)\beta, \left(\Pi_m - \widehat{\Pi}_m\right)\beta\right\rangle + \left\|\left(\Pi_m - \widehat{\Pi}_m\right)\beta\right\|_{\Gamma}^2$$

and that it is enough to focus on the first term and to prove the bound for

$$\mathbb{E}\left[\left\|\left(\Gamma_n - \Gamma\right)\left(\Pi_m - \widehat{\Pi}_m\right)\beta\right\|\right],$$

after a Cauchy-Schwartz's Inequality coupled with the fact that $\left\|\left(\Pi_m - \widehat{\Pi}_m\right)\beta\right\| \leq 2\|\beta\|$.

Now by Proposition 3

$$\left(\Gamma_n - \Gamma\right)\left(\Pi_m - \widehat{\Pi}_m\right)\beta \mathbf{1}_{\mathcal{A}_n} = \frac{1}{2\pi i}\sum_{j=1}^{D_m}\int_{\partial\Omega_j}(\Gamma_n - \Gamma)R^{1/2}(\zeta)[I - T(\zeta)]^{-1}T(\zeta)R^{1/2}(\zeta)\beta d\zeta \mathbf{1}_{\mathcal{A}_n}.$$

Hence, by definition of $\mathcal{A}_n$, $\|(I-T(\zeta))^{-1}\|_\infty \mathbf{1}_{\mathcal{A}_n} \leq 2$ and $\|T(\zeta)\|_\infty \mathbf{1}_{\mathcal{A}_n} \leq \frac{\mathbf{a}_j \ln n}{\sqrt{n}}$, then

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\Gamma_n-\Gamma)\left(\Pi_m-\widehat{\Pi}_m\right)\beta\right\|\mathbf{1}_{\mathcal{A}_n}\right] \\
&\qquad\leq \frac{1}{\pi}\sum_{j=1}^{D_m}\int_{\partial\Omega_j}\mathbb{E}\left[\left\|(\Gamma_n-\Gamma)R^{1/2}(\zeta)\right\|_\infty \|T(\zeta)\|_\infty \mathbf{1}_{\mathcal{A}_n}\right]\left\|R^{1/2}(\zeta)\beta\right\| d\zeta \\
&\qquad\leq \frac{\ln n}{\pi\sqrt{n}}\sum_{j=1}^{D_m}\mathbf{a}_j\int_{\partial\Omega_j}\left\|R^{1/2}(\zeta)P_r\right\|_\infty\left\|\beta^\Diamond\right\|\mathbb{E}\left[\left\|(\Gamma_n-\Gamma)R^{1/2}(\zeta)\right\|_\infty\mathbf{1}_{\mathcal{A}_n}\right]d\zeta.
\end{aligned}$$

Recall that, by equations (A.12) and (A.13)

$$\left\|R^{1/2}(\zeta)P_r\right\|_\infty \leq \begin{cases} C_1\mathbf{a}_j^{1/2}j^{(a-r)_+/2} & \text{in the polynomial case } \mathbf{(P)} \\ C_2\mathbf{a}_j^{1/2}j^{-r/2}e^{j^a/2} & \text{in the exponential case } \mathbf{(E)}\end{cases}$$

(the definitions of $\beta^\diamond$ and P_r are given in the proof of Lemma 18). Then, in the polynomial case **(P)**,

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\Gamma_n-\Gamma)\left(\Pi_m-\widehat{\Pi}_m\right)\beta\right\|\mathbf{1}_{\mathcal{A}_n}\right] \\
&\qquad\leq \frac{\ln n}{\pi\sqrt{n}}\left\|\beta^\Diamond\right\|\sum_{j=1}^{D_m}\mathbf{a}_j^{3/2}j^{(a-r)_+/2}\int_{\partial\Omega_j}\sqrt{\mathbb{E}\left[\left\|(\Gamma_n-\Gamma)R^{1/2}(\zeta)\right\|_{HS}^2\mathbf{1}_{\mathcal{A}_n}\right]}d\zeta
\end{aligned} \tag{A.15}$$

and in the exponential case **(E)**

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\Gamma_n-\Gamma)\left(\Pi_m-\widehat{\Pi}_m\right)\beta\right\|\mathbf{1}_{\mathcal{A}_n}\right] \\
&\qquad\leq \frac{\ln n}{\pi\sqrt{n}}\left\|\beta^\Diamond\right\|\sum_{j=1}^{D_m}\mathbf{a}_j^{3/2}j^{-r/2}e^{j^a/2}\int_{\partial\Omega_j}\sqrt{\mathbb{E}\left[\left\|(\Gamma_n-\Gamma)R^{1/2}(\zeta)\right\|_{HS}^2\mathbf{1}_{\mathcal{A}_n}\right]}d\zeta.
\end{aligned} \tag{A.16}$$

Treating $\mathbb{E}\left[\left\|(\Gamma_n-\Gamma)R^{1/2}(\zeta)\right\|_{HS}^2\right]$ with computations similar to those carried previously we get, for all $\zeta\in\partial\Omega_j$

$$\mathbb{E}\left[\left\|(\Gamma_n-\Gamma)R^{1/2}(\zeta)\right\|_{HS}^2\right] \leq \frac{C'\mathbf{a}_j}{n},$$

with $C'=\mathrm{Tr}(\Gamma)\max\{1,b-1\}$. Putting into Equation (A.15) we obtain

$$\mathbb{E}\left[\left\|(\Gamma_n-\Gamma)\left(\Pi_m-\widehat{\Pi}_m\right)\beta\right\|\mathbf{1}_{\mathcal{A}_n}\right] \leq \frac{C'\ln n}{\pi n}\left\|\beta^\Diamond\right\|\sum_{j=1}^{D_m}\mathbf{a}_j^2 j^{(a-r)_+/2}\pi\delta_j,$$

for the polynomial case **(P)** and, from Equation (A.16)

$$\mathbb{E}\left[\left\|(\Gamma_n-\Gamma)\left(\Pi_m-\widehat{\Pi}_m\right)\beta\right\|\mathbf{1}_{\mathcal{A}_n}\right] \leq \frac{C'\ln n}{\pi n}\left\|\beta^\Diamond\right\|\sum_{j=1}^{D_m}\mathbf{a}_j^2 j^{-r/2}e^{j^a/2}\pi\delta_j,$$

for the exponential case **(E)**.

Considering again two cases related to the rate of decrease for the eigenvalues we see first that for an exponential decay the term above is bounded up to a constant by $(\ln n)/n$. since $\delta_j \leq \lambda_j \leq C_E e^{-j^a}$. Secondly in case of polynomial decay we get :

$$\mathbb{E}\left[\left\|(\Gamma_n - \Gamma)\left(\Pi_m - \widehat{\Pi}_m\right)\beta\right\| \mathbf{1}_{\mathcal{A}_n}\right] \leq \frac{C' \ln n}{n} \left\|\beta^{\Diamond}\right\| \sum_{j=1}^{D_m} j^2 j^{-r/2} \delta_j \ln^{3/2}(j+1) \leq C\frac{D_m}{n},$$

using the fact that $\delta_j \leq \lambda_j \leq C_P j^{-a}$ and that $a + r/2 > 2$.

Thus the proof is finished by Lemma 16:

$$\mathbb{E}\left[\left\|\left(\Pi_m - \widehat{\Pi}_m\right)\beta\right\|_{\Gamma_n}^2 \mathbf{1}_{\mathcal{A}_n^{\complement}}\right] \leq \|\beta\|_{\Gamma}^2 \mathbb{P}(\mathcal{A}_n^{\complement}) \leq \frac{C''(b,\Gamma)}{n^2}.$$

□

A.5 Upper-bound on $\mathbb{P}\left(\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}}\right)$

The aim of this section is to bound the probability of $\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}}$.

Lemma 20. *Under assumptions **H2**, **H3** and **H4** and if the decreasing rate of $(\lambda_j)_{j\geq 1}$ is given by **(P)** or **(E)**,*

$$\mathbb{P}\left(\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \cup \widehat{\Delta}^{\complement}_{N_n}\right) \leq C/n^6,$$

with $C > 0$ independent of n and we recall that, for all $m \geq 1$,

$$\widehat{\Delta}_m := \left\{\forall f \in \widehat{S}_m, \|f\|_{\Gamma}^2 \leq \rho_0 \|f\|_{\Gamma_n}^2\right\}.$$

Proof. First remark that

$$\begin{aligned}\mathbb{P}\left(\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \cup \widehat{\Delta}^{\complement}_{N_n}\right) &= \mathbb{P}\left(\left\{\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \cup \widehat{\Delta}^{\complement}_{N_n}\right\} \cap \{\widehat{N}_n^{(FPCR)} \leq N_n\}\right) \\ &\quad + \mathbb{P}\left(\left\{\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \cup \widehat{\Delta}^{\complement}_{N_n}\right\} \cap \{\widehat{N}_n^{(FPCR)} > N_n\}\right).\end{aligned}$$

Now, if $\widehat{N}_n^{(FPCR)} \leq N_n$, then $\widehat{S}_{\widehat{N}_n^{(FPCR)}} \subset \widehat{S}_{N_n}$ and $\widehat{\Delta}_{N_n} \subset \widehat{\Delta}_{\widehat{N}_n^{(FPCR)}}$, hence

$$\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \subset \widehat{\Delta}^{\complement}_{N_n}.$$

Then

$$\begin{aligned}\mathbb{P}\left(\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \cup \widehat{\Delta}^{\complement}_{N_n}\right) &\leq \mathbb{P}\left(\widehat{\Delta}^{\complement}_{N_n} \cap \{\widehat{N}_n^{(FPCR)} \leq N_n\}\right) + \mathbb{P}\left(\widehat{N}_n^{(FPCR)} > N_n\right) \\ &\leq \mathbb{P}\left(\widehat{\Delta}^{\complement}_{N_n}\right) + \mathbb{P}\left(\widehat{N}_n^{(FPCR)} > N_n\right).\end{aligned}$$

The second term of this equality is bounded by Cn^{-6} by Lemma 21. It remains to bound

the first term which can be written the following way:

$$\widehat{\Delta}^{\complement}_{N_n} = \left\{ \inf_{f\in\widehat{S}_{N_n}} \frac{\|f\|^2_{\Gamma_n}}{\|f\|^2_{\Gamma}} < \rho_0^{-1} \right\}.$$

Let $f = \sum_{j=1}^{D_{N_n}} \alpha_j \widehat{\psi}_j \in \widehat{S}_{N_n}$, we have a.s.

$$\|f\|^2_{\Gamma_n} = \sum_{j=1}^{D_{N_n}} \widehat{\lambda}_j \alpha_j^2 = |\widehat{\Lambda}_n \alpha|_2^2,$$

where $|\cdot|_2$ is the norm of $\mathbb{R}^{D_{N_n}}$ defined by $|x|_2^2 = \sum_{j=1}^{D_{N_n}} x_i^2$ for all $x = (x_1, ..., x_{D_{N_n}})' \in \mathbb{R}^{D_{N_n}}$ and $\widehat{\Lambda}_n$ is the diagonal matrix with entries $\left\{\sqrt{\widehat{\lambda}_1}, ..., \sqrt{\widehat{\lambda}_{D_{N_n}}}\right\}$.
Moreover

$$\|f\|^2_{\Gamma} = \sum_{j,k=1}^{D_{N_n}} \alpha_j \alpha_k \langle \Gamma^{1/2}\widehat{\psi}_j, \Gamma^{1/2}\widehat{\psi}_k \rangle = \alpha' \Psi_n \alpha,$$

where Ψ_n is the symmetric and positive-definite matrix

$$\Psi_n := \left(\langle \Gamma^{1/2}\widehat{\psi}_j, \Gamma^{1/2}\widehat{\psi}_k \rangle \right)_{1\le j,k\le D_{N_n}}.$$

Then,

$$\frac{\|f\|^2_{\Gamma_n}}{\|f\|^2_{\Gamma}} = \frac{|\widehat{\Lambda}_n\alpha|_2^2}{\alpha'\Psi_n\alpha} = \frac{|\widehat{\Lambda}_n\alpha|_2^2}{|\Psi_n^{1/2}\alpha|_2^2} = \frac{|\widehat{\Lambda}_n\Psi_n^{-1/2}\Psi_n^{1/2}\alpha|_2^2}{|\Psi_n^{1/2}\alpha|_2^2}.$$

Now

$$\inf_{f\in\widehat{S}_{N_n}} \frac{\|f\|^2_{\Gamma_n}}{\|f\|^2_{\Gamma}} = \inf_{\alpha\in\mathbb{R}^{N_n}\setminus\{0\}} \frac{\alpha'\Psi_n^{-1/2}\widehat{\Lambda}_n^2\Psi_n^{-1/2}\alpha}{|\alpha|_2^2} = \min\{\lambda, \lambda \text{ eigenvalue of } \Psi_n^{-1/2}\widehat{\Lambda}_n^2\Psi_n^{-1/2}\}.$$

On the set $\mathcal{A}_n$ we have for all, $j = 1, ..., N_n$, $\widehat{\lambda}_j > 0$. Hence the matrix $\widehat{\Lambda}_n$ is invertible, therefore

$$\inf_{f\in\widehat{\mathcal{S}}_n} \frac{\|f\|^2_{\Gamma_n}}{\|f\|^2_{\Gamma}} = \rho\left((\Psi_n^{-1/2}\widehat{\Lambda}_n^2\Psi_n^{-1/2})^{-1}\right)^{-1} = \rho(\widehat{\Lambda}_n^{-1}\Psi_n\widehat{\Lambda}_n^{-1})^{-1},$$

where, for a matrix A, $\rho(A) = \max\{|\lambda|, \lambda \text{ is a complex eigenvalue of } A\}$ denotes the spectral radius of A. We have then

$$\mathbb{P}\left(\widehat{\Delta}^{\complement}_{\widehat{N}_n^{(FPCR)}} \cap \left\{\widehat{N_n} < N_n\right\}\right) \le \mathbb{P}(\mathcal{A}_n \cap \{\rho(\widehat{\Lambda}_n^{-1}\Psi_n\widehat{\Lambda}_n^{-1}) > \rho_0\}) + \mathbb{P}(\mathcal{A}_n^{\complement}). \tag{A.17}$$

By Lemma 16 in Section A.2 (p. 165), $\mathbb{P}(\mathcal{A}_n{}^{\complement}) \le C/n^6$, with C depending only on Γ and b. Thus it remains to control the spectral radius of $\widehat{\Lambda}_n^{-1}\Psi_n\widehat{\Lambda}_n^{-1}$.

We define a linear (random) application O from $\mathbb{R}^{D_{N_n}}$ to $\mathbb{H}$ by:

$$O : \alpha = (\alpha_1, ..., \alpha_{D_{N_n}})' \mapsto \sum_{j=1}^{D_{N_n}} \alpha_j \widehat{\psi}_j.$$

We denote by O^* the adjoint of O, which is the linear map from $\mathbb{H}$ to $\mathbb{R}^{D_{N_n}}$ defined by:

$$O^* : f \mapsto (\langle f, \widehat{\psi}_j \rangle_{1 \leq j \leq D_{N_n}}).$$

We can check that $OO^* = \widehat{\Pi}_{D_{N_n}}$ and Ψ_n is the matrix of the linear map $O^*\Gamma O$ in the standard basis of $\mathbb{R}^{N_n}$.

It is known that the spectral radius of an operator is equal to the spectral radius of its adjoint, then,

$$\rho(\widehat{\Lambda}_n^{-1}\Psi_n\widehat{\Lambda}_n^{-1}) = \rho(\widehat{\mathcal{L}}_n^{-1}O^*\Gamma O\widehat{\mathcal{L}}_n^{-1}) = \rho(\Gamma^{1/2}O\widehat{\mathcal{L}}_n^{-1}\widehat{\mathcal{L}}_n^{-1}O^*\Gamma^{1/2}), \tag{A.18}$$

where $\widehat{\mathcal{L}}_n$ denotes the linear endormorphism of $\mathbb{R}^{D_{N_n}}$ whose matrix in the standard basis is $\widehat{\Lambda}_n$. Denote by Π_{N_n} the orthogonal projector onto $S_{N_n} = \text{span}\{\psi_1, ..., \psi_{D_{N_n}}\}$. Moreover, let $\Gamma^\dagger$ (resp. $\Gamma_n^\dagger$) the pseudo-inverse of operator Γ (resp. Γ_n) on S_{N_n} (resp. $\widehat{\mathcal{S}}_n$), defined by

$$\Gamma^\dagger f := \sum_{j=1}^{D_{Nn}} \frac{\langle f, \psi_j \rangle}{\lambda_j}\psi_j \text{ and } \Gamma_n^\dagger f := \sum_{j=1}^{D_{Nn}} \frac{\langle f, \widehat{\psi}_j \rangle}{\widehat{\lambda}_j}\widehat{\psi}_j \mathbf{1}_{\{\widehat{\lambda}_j > 0\}}. \tag{A.19}$$

We have $\Gamma^{1/2}\Gamma^\dagger\Gamma^{1/2} = \Pi_{N_n}$ and $O\widehat{\mathcal{L}}_n^{-2}O^* = \Gamma_n^\dagger$, then,

$$\Gamma^{1/2}O\widehat{\mathcal{L}}_n^{-1}\widehat{\mathcal{L}}_n^{-1}O^*\Gamma^{1/2} = \Gamma^{1/2}\Gamma_n^\dagger\Gamma^{1/2} = \Gamma^{1/2}(\Gamma^\dagger + \Gamma_n^\dagger - \Gamma^\dagger)\Gamma^{1/2} = \Pi_{N_n} + \Gamma^{1/2}(\Gamma_n^\dagger - \Gamma^\dagger)\Gamma^{1/2},$$

and by Equation (A.18)

$$\rho(\widehat{\Lambda}_n^{-1}\Psi_n\widehat{\Lambda}_n^{-1}) = \|\Pi_{N_n} + \Gamma^{1/2}(\Gamma_n^\dagger - \Gamma^\dagger)\Gamma^{1/2}\|_\infty \leq 1 + \|\Gamma^{1/2}(\Gamma_n^\dagger - \Gamma^\dagger)\Gamma^{1/2}\|_\infty,$$

where $\|\cdot\|_\infty$ denotes the usual operator norm.
Now

$$\mathbb{P}(\mathcal{A}_n \cap \{\rho(\widehat{\Lambda}_n^{-1}\Psi_n\widehat{\Lambda}_n^{-1}) > \rho_0\}) \leq \mathbb{P}\left(\mathcal{A}_n \cap \left\{\|\Gamma^{1/2}(\Gamma_n^\dagger - \Gamma^\dagger)\Gamma^{1/2}\|_\infty > \rho_0 - 1\right\}\right).$$

Let γ be the contour defined by Figure A.1 with $m = N_n$.

We have by Proposition 3 and the fact that $(I-T(z))^{-1}T(z) = T(z)+(I-T(z))^{-1}T^2(z)$,

$$\Gamma^{1/2}\left[\Gamma_n^{\dagger}-\Gamma^{\dagger}\right]\Gamma^{1/2}\mathbf{1}_{\mathcal{A}_n} = \mathbf{1}_{\mathcal{A}_n}\frac{1}{2i\pi}\int_{\gamma}\frac{1}{z}\Gamma^{1/2}R^{1/2}(z)\left[I-T(z)\right]^{-1}T(z)R^{1/2}(z)\Gamma^{1/2}dz \tag{A.20}$$

$$= \mathbf{1}_{\mathcal{A}_n}\frac{1}{2i\pi}\int_{\gamma}\frac{1}{z}\Gamma^{1/2}R(z)\left(\Gamma_n-\Gamma\right)R(z)\Gamma^{1/2}dz + \mathbf{1}_{\mathcal{A}_n}\frac{1}{2i\pi}\int_{\gamma}\frac{1}{z}\Gamma^{1/2}R^{1/2}(z)\left[I-T(z)\right]^{-1}T^2(z)R^{1/2}(z)\Gamma^{1/2}dz. \tag{A.21}$$

Now, we consider separately the two decreasing rates of the λ_j's.

Exponential decrease : $c\exp(-j^a) \leq \lambda_j \leq C\exp(-j^a)$, with $a > 0$. By Equation (A.20) and the fact that $\|(I-T(z))^{-1}\|_{\infty} < 2$, on the set $\mathcal{A}_n$

$$\|\Gamma^{1/2}(\Gamma_n^{\dagger}-\Gamma^{\dagger})\Gamma^{1/2}\|_{\infty}\mathbf{1}_{\mathcal{A}_n} \leq \frac{1}{2\pi}\left\|\int_{\gamma}\frac{1}{z}\Gamma^{1/2}R^{1/2}(z)\left[I-T(z)\right]^{-1}T(z)R^{1/2}(z)\Gamma^{1/2}dz\right\|_{\infty}\mathbf{1}_{\mathcal{A}_n}$$
$$\leq \pi^{-1}\sup_{z\in\gamma}\left[\|T(z)\|_{\infty}\right]\int_{\gamma}\frac{1}{|z|}\left\|\Gamma^{1/2}R^{1/2}(z)\right\|_{\infty}^2 dz\mathbf{1}_{\mathcal{A}_n}.$$

For $z \in \mathrm{supp}(\gamma)$, the eigenvalues of the operator $\Gamma^{1/2}R^{1/2}(z)$ are $\left\{\lambda_j^{1/2}(z-\lambda_j)^{-1/2}, j \geq 1\right\}$, then

$$\left\|\Gamma^{1/2}R^{1/2}(z)\right\|_{\infty}^2 = \sup_{j\geq 1}\left\{\frac{\lambda_j}{|z-\lambda_j|}\right\},$$

and

$$\left|\int_{\gamma}\frac{1}{|z|}\left\|\Gamma^{1/2}R^{1/2}(z)\right\|_{\infty}^2 dz\right| \leq C + 2\int_0^{2\lambda_1/\delta_{D_{N_n}}}\frac{du}{1+u^2} \leq C',$$

where C and C' are independent of n and the last inequality comes from the fact that in the exponential case, there exists a constant $c > 0$ such that $\delta_{D_{N_n}}/\lambda_{D_{N_n}} \geq c$. Then by definition of $\mathcal{A}_n$,

$$\mathbb{P}\left(\mathcal{A}_n \cap \left\{\|\Gamma^{1/2}(\Gamma_n^{\dagger}-\Gamma^{\dagger})\Gamma^{1/2}\|_{\infty} > \rho_0 - 1\right\}\right) \leq \mathbb{P}\left(C'\sup_{z\in\mathrm{supp}(\gamma)}\left[\|T(z)\|_{\infty}\right] > \pi(\rho_0-1)\right)$$
$$= 0 \text{ if } n \text{ is sufficiently large.}$$

The result comes from the fact that $D_{N_n} \leq 20\sqrt{n/\ln^3 n}$.

Polynomial decrease : $cj^{-a} \leq \lambda_j \leq Cj^{-a}$, $a > 1$. Denote by T_1 and T_2 the two terms of Equation (A.21) i.e.

$$\begin{aligned} T_1 &= \mathbf{1}_{\mathcal{A}_n} \frac{1}{2i\pi} \int_\gamma \frac{1}{z} \Gamma^{1/2} R(z) \left(\Gamma_n - \Gamma\right) R(z) \Gamma^{1/2} dz, \\ T_2 &= \mathbf{1}_{\mathcal{A}_n} \frac{1}{2i\pi} \int_\gamma \frac{1}{z} \Gamma^{1/2} R^{1/2}(z) \left[I - T(z)\right]^{-1} T^2(z) R^{1/2}(z) \Gamma^{1/2} dz. \end{aligned}$$

First we control T_2, the proof in the exponential case leads us to:

$$\|T_2\|_\infty \leq \pi^{-1} \sup_{z\in\gamma} \left[\|T(z)\|_\infty^2\right] \int_\gamma \frac{1}{|z|} \left\|\Gamma^{1/2} R^{1/2}(z)\right\|_\infty^2 dz,$$

and

$$\begin{aligned} \int_\gamma \frac{1}{|z|} \left\|\Gamma^{1/2} R^{1/2}(z)\right\|_\infty^2 dz &\leq C + \int_0^{2\lambda_1/\delta_{D_{N_n}}} \frac{\lambda_{D_{N_n}} du}{\sqrt{(\lambda_{D_{N_n}} - \delta_{D_{N_n}})^2 + \delta_{D_{N_n}}^2 u^2}\sqrt{1+u^2}} \\ &\leq C + \int_0^1 \frac{\lambda_{D_{N_n}}}{\lambda_{D_{N_n}} - \delta_{D_{N_n}}} du + \int_1^{2\lambda_1/\delta_{D_{N_n}}} \frac{\lambda_{D_{N_n}} du}{\sqrt{(\lambda_{D_{N_n}} - \delta_{D_{N_n}})^2 + \delta_{D_{N_n}}^2 u^2}\sqrt{1+u^2}} \\ &\leq 1 + C + \int_1^{2\lambda_1/\delta_{D_{N_n}}} \frac{du}{\sqrt{1+u^2}} \leq C' \ln(D_{N_n}), \end{aligned}$$

with $C, C' > 0$ independent of n. Then by definition of $\mathcal{A}_n$

$$\begin{aligned} \mathbb{P}(\|T_2\|_\infty > (\rho_0 - 1)/2) &\leq \mathbb{P}\left(C' \ln(D_{N_n}) \sup_{z\in\gamma} \left[\|T(z)\|_\infty^2\right] > \pi(\rho_0 - 1)/2\right) \\ &= 0 \text{ if } n \text{ is sufficiently large.} \end{aligned}$$

Now, we can calculate explicitly the term T_1:

$$T_1 = \mathbf{1}_{\mathcal{A}_n} \frac{1}{2i\pi} \int_\gamma \sum_{j,k\geq 1} \frac{\sqrt{\lambda_k}\sqrt{\lambda_j}}{z(z-\lambda_k)(z-\lambda_j)} \pi_k (\Gamma_n - \Gamma) \pi_j dz.$$

By the Residue Theorem

$$\frac{1}{2i\pi} \int_\gamma \frac{dz}{z(z-\lambda_k)(z-\lambda_j)} = \begin{cases} -\frac{1}{\lambda_j \lambda_k} & \text{if } j \neq k, j \leq D_{N_n} \text{ and } k \leq D_{N_n} \\ \frac{1}{\lambda_j(\lambda_j - \lambda_k)} & \text{if } j \leq D_{N_n} < k \\ \frac{1}{\lambda_k(\lambda_k - \lambda_j)} & \text{if } k \leq D_{N_n} < j \\ 0 & \text{otherwise} \end{cases}.$$

Then

$$\begin{aligned} T_1 &= -\mathbf{1}_{\mathcal{J}_n} \sum_{\substack{j,k=1 \\ j\neq k}}^{D_{N_n}} \frac{1}{\sqrt{\lambda_j\lambda_k}} \pi_k(\Gamma_n - \Gamma)\pi_j + \sum_{j=1}^{D_{N_n}} \sum_{k>D_{N_n}} \frac{\sqrt{\lambda_k}}{\sqrt{\lambda_j}(\lambda_j - \lambda_k)} \pi_k(\Gamma_n - \Gamma)\pi_j \\ &\quad + \sum_{j>D_{N_n}} \sum_{k=1}^{D_{N_n}} \frac{\sqrt{\lambda_j}}{\sqrt{\lambda_k}(\lambda_k - \lambda_j)} \pi_k(\Gamma_n - \Gamma)\pi_j \\ &= -\mathbf{1}_{\mathcal{J}_n} \sum_{\substack{j,k=1 \\ j\neq k}}^{D_{N_n}} \frac{\pi_k}{\sqrt{\lambda_k}} \Gamma_n \frac{\pi_j}{\sqrt{\lambda_j}} + \sum_{j=1}^{D_{N_n}} \left(s_j \Gamma_n \frac{\pi_j}{\sqrt{\lambda_j}} + \frac{\pi_j}{\sqrt{\lambda_j}} \Gamma_n s_j \right) = T_1' + T_1'', \end{aligned}$$

where $s_j := \sum_{k>D_{N_n}} \frac{\sqrt{\lambda_k}}{\lambda_j - \lambda_k} \pi_k$. The term Γ disappears because $\pi_j \Gamma \pi_k = 0$ if $j \neq k$.

We control separately the operators T_1' and T_1''. We have:

$$\|T_1'\|_\infty^2 \leq \sum_{p,q\geq 1} \langle T_1' \psi_p, \psi_q \rangle^2 = \sum_{\substack{p,q=1 \\ p\neq q}}^{D_{N_n}} \left(\frac{1}{n} \sum_{i=1}^{n} \xi_p^{(i)} \xi_q^{(i)} \right)^2,$$

where we recall that $\xi_p^{(i)} = \langle X_i, \psi_p \rangle / \sqrt{\lambda_p}$. Then

$$\begin{aligned} \mathbb{P}(\|T_1'\|_\infty > (\rho_0 - 1)/4) &\leq \mathbb{P}\left(\sum_{\substack{p,q=1 \\ p\neq q}}^{D_{N_n}} \left(\frac{1}{n} \sum_{i=1}^{n} \xi_p^{(i)} \xi_q^{(i)} \right)^2 > \frac{(\rho_0 - 1)^2}{16} \right) \\ &\leq \sum_{\substack{p,q=1 \\ p\neq q}}^{D_{N_n}} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \xi_p^{(i)} \xi_q^{(i)} \right| > \frac{\rho_0 - 1}{4 D_{N_n}} \right). \end{aligned}$$

For all $p \neq q$, the sequence of random variables $\left\{ \xi_p^{(i)} \xi_q^{(i)}, i = 1, ..., n \right\}$ is independent and centred and by assumptions **H2** and **H3**,

$$\mathbb{E}\left[|\xi_p \xi_q|^m \right] \leq m! b^{m-1},$$

then Lemma 23 and the condition $D_{N_n} \leq 20\sqrt{n/\ln^3 n}$ implies

$$\mathbb{P}\left(\|T_1'\|_\infty > \frac{\rho_0 - 1}{4} \right) \leq 2 {D_{N_n}}^2 \exp\left(-C_1' \frac{n}{{D_{N_n}}^2} \right) \leq C_2' n^{-6},$$

with $C_1' = \frac{(\rho_0 - 1)^2}{32(2b + (\rho_0 - 1)/4)}$ and C_2' depends only on b and ρ_0.

We deal now with the operator T_1'', we can rewrite it as an array of independent random variables with values in $\mathcal{H}$, the set of the Hilbert-Schmidt operators of $\mathbb{H}$ equipped with the

usual norm $\|T\|_{HS}^2 = \sum_{p,q\geq 1}\langle T\psi_p, \psi_q\rangle^2$, i.e.

$$T_1'' = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{D_{N_n}}\left(s_jX_i\otimes\frac{\pi_j}{\sqrt{\lambda_j}}X_i + \frac{\pi_j}{\sqrt{\lambda_j}}X_i\otimes s_jX_i\right) = \frac{1}{n}\sum_{i=1}^{n}Z_i,$$

where, for all $f, g \in \mathbb{H}$, the operator $f\otimes g$ is defined by $f\otimes g : h \mapsto \langle f, h\rangle g$. In order to apply the exponential inequality for centred Hilbert valued random variable recalled in Appendix B (Lemma 23, p. 180), we have to find two constants B and c such that

$$\mathbb{E}\left[\|Z_i\|_{HS}^m\right] \leq (m!/2)B^2c^{m-2}.$$

We compute first $\|Z_i\|_{HS}^2$

$$\|Z_i\|_{HS}^2 \leq 2\sum_{p,q\geq 1}\langle\sum_{j=1}^{D_{N_n}}\left(\frac{\pi_j}{\sqrt{\lambda_j}}X_i\otimes s_jX_i\right)\psi_p,\psi_q\rangle^2 = \sum_{p=1}^{D_{N_n}}\sum_{q>D_{N_n}}\frac{\lambda_q^2}{(\lambda_p-\lambda_q)^2}\left(\xi_p^{(i)}\xi_q^{(i)}\right)^2.$$

Now, by assumptions **H2** and **H3**,

$$\begin{aligned}\mathbb{E}\left[\|Z_i\|_{HS}^m\right] &\leq \mathbb{E}\left[\|Z_i\|_{HS}^{2m}\right]^{1/2}\\ &= \left(\sum_{p_1,\ldots,p_m=1}^{D_{N_n}}\sum_{q_1,\ldots,q_m>D_{N_n}}\prod_{j=1}^{m}\frac{\lambda_{q_j}^2}{(\lambda_{p_j}-\lambda_{q_j})^2}\mathbb{E}\left[\left(\xi_{p_1}^{(i)}\ldots\xi_{p_m}^{(i)}\right)^2\right]\mathbb{E}\left[\left(\xi_{q_1}^{(i)}\ldots\xi_{q_m}^{(i)}\right)^2\right]\right)^{1/2}\\ &\leq m!b^m\left(\sum_{p=1}^{D_{N_n}}\sum_{q>D_{N_n}}\frac{\lambda_q^2}{(\lambda_p-\lambda_q)^2}\right)^{m/2}.\end{aligned}$$

We apply then Lemma 23 with $B^2 = 2b^2\sum_{p=1}^{D_{N_n}}\sum_{q>D_{N_n}}\frac{\lambda_q^2}{(\lambda_p-\lambda_q)^2}$ and $c = b\left(\sum_{p=1}^{D_{N_n}}\sum_{q>D_{N_n}}\frac{\lambda_q^2}{(\lambda_p-\lambda_q)^2}\right)^{1/2}$ and obtain, with the condition $D_{N_n}\leq 20\sqrt{n/\ln^3 n}$,

$$\mathbb{P}\left(\|T_1''\|_\infty > \frac{\rho_0-1}{4}\right)\leq\mathbb{P}\left(\|T_1''\|_{HS} > \frac{\rho_0-1}{4}\right)\leq 2\exp\left(-C_1''\frac{n}{D_{N_n}}\right)\leq C_2''n^{-6},$$

where $C_1'' := (\rho_0-1)^2/(2\delta b^2 + \sqrt{2\delta}b(\rho_0-1)/4)$, C_2'' depends only on ρ_0 and δ where $\delta>0$ depends only on the sequence $(\lambda_j)_{j\geq 1}$ and verifies, for all $p\leq D_{N_n}$,

$$\sum_{q>D_{N_n}}\frac{\lambda_q^2}{(\lambda_p-\lambda_q)^2}\leq D_{N_n}\delta/2.$$

Then the proof is finished with the results of Lemma 21 below. □

Lemma 21. *If Assumption **H2** is fulfilled and $n\geq 6$, then*

$$\mathbb{P}(\widehat{N}_n^{(FPCR)} > N_n)\leq Cn^{-6},$$

with C independent of β and n.

Proof. The sequence $(\lambda_j)_{j\geq 1}$ being non-increasing we have

$$\mathbb{P}(\widehat{N}_n^{(FPCR)} > N_n) \leq \mathbb{P}(\lambda_{N_n+1} \geq \lambda_{D_{\widehat{N}_n^{(FPCR)}}}) \leq \mathbb{P}\left(\left\{\lambda_{D_{N_n}+1} \geq \lambda_{D_{\widehat{N}_n^{(FPCR)}}}\right\} \cap \mathcal{A}_n\right) + \mathbb{P}\left(\mathcal{A}_n^{\complement}\right),$$

where $\mathcal{A}_n$ is defined by Equation (A.7) in Section A.2. Then by definition of $\mathcal{A}_n$,

$$\left\{\lambda_{D_{N_n}+1} \geq \lambda_{D_{\widehat{N}_n^{(FPCR)}}}\right\} \cap \mathcal{A}_n \subset \left\{\lambda_{D_{N_n}+1} \geq \widehat{\lambda}_{D_{\widehat{N}_n^{(FPCR)}}} - \frac{\delta_{D_{\widehat{N}_n^{(FPCR)}}}}{2}\right\} = \emptyset,$$

since $\lambda_{D_{N_n}+1} < n^{-2}$, $\lambda_{D_{\widehat{\lambda}_{N_n}}} \geq \mathfrak{s}_n$ and $\delta_{D_{\widehat{N}_n^{(FPCR)}}} \leq \frac{\lambda_{D_{\widehat{N}_n^{(FPCR)}}}}{4} \leq \frac{n^{-2}}{4}$. Thus the proof is finished by Lemma 16. □

Annexe B

Inégalités de concentration

La majorité des preuves présentées dans cette thèse est basée sur des inégalités de concentration. Nous présentons dans cette section les énoncés de ces résultats.

B.1 Bernstein's Inequality

B.1.1 For real random variables

Lemma 22. (Birgé and Massart, 1998, Lemma 8) *Let $T_1, T_2, \ldots, T_n$ be independent random variables and $S_n(T) = \sum_{i=1}^n (T_i - \mathbb{E}[T_i])$. Assume that, there exist $v > 0$ and $b_0 > 0$ such that*

$$Var(T_1) \leq v^2 \text{ and } \forall \ell \geq 2, \ \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[|T_i|^\ell\right] \leq \frac{\ell!}{2} v^2 b_0^{\ell-2}.$$

Then, for $\eta > 0$,

$$\begin{aligned} \mathbb{P}\left(\left|\frac{1}{n}S_n(T)\right| \geq \eta\right) &\leq 2\exp\left(-\frac{n\eta^2/2}{v^2 + b_0\eta}\right), \\ &\leq 2\min\left\{\exp\left(-\frac{n\eta^2}{4v^2}\right), \exp\left(-\frac{n\eta}{4b_0}\right)\right\}. \end{aligned} \tag{B.1}$$

B.1.2 For Hilbert-valued random variables

Lemma 23 (Bosq (2000))**.** *Let $X_1, \ldots, X_n$ be centred independent random variables lying in a separable Hilbert space $\mathbb{H}$. If, for some constants v and b_0, we have for all $\ell \geq 2$,*

$$\sum_{i=1}^n \mathbb{E}\left[\|X_i\|^\ell\right] \leq \frac{\ell!}{2} v^2 b_0^{\ell-2}.$$

Then, for all $\eta > 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| \geq \eta\right) \leq 2\exp\left(-\frac{\eta^2}{2v^2 + 2b_0\eta}\right).$$

B.2 Control of linear empirical processes: Talagrand's Inequality and some corollaries

Lemma 24. *[Talagrand's Inequality] Let $\xi_1, \ldots, \xi_n$ be i.i.d. random variables, and define $\nu_n(r) = \frac{1}{n}\sum_{i=1}^n r(\xi_i) - \mathbb{E}[r(\xi_i)]$, for r belonging to a countable class $\mathcal{R}$ of real-valued mea-*

surable functions. Then, for $\delta > 0$, there exists a universal constant C such that

$$\mathbb{E}\left[\left(\sup_{r\in\mathcal{R}}\left(\nu_n\left(r\right)\right)^2 - c(\delta)(H^\nu)^2\right)_+\right] \leq C\left\{\frac{v^\nu}{n}\exp\left(-\frac{\delta}{6}\frac{n(H^\nu)^2}{v^\nu}\right) + \frac{(M_1^\nu)^2}{C^2(\delta)n^2}\exp\left(-\frac{1}{21\sqrt{2}}C(\delta)\sqrt{\delta}\frac{nH^\nu}{M_1^\nu}\right)\right\},$$

with, $C(\delta) = (\sqrt{1+\delta} - 1) \wedge 1$, $c(\delta) = 2(1+2\delta)$ and

$$\sup_{r\in\mathcal{R}}\|r\|_{L^\infty} \leq M_1^\nu,\ \mathbb{E}\left[\sup_{r\in\mathcal{R}}|\nu_n(r)|\right] \leq H^\nu,\ \textit{and}\ \sup_{r\in\mathcal{R}} Var\left(r\left(\xi_1\right)\right) \leq v^\nu.$$

Lemma 24 above is a classical consequence of the Talagrand Inequality given in Klein and Rio (2005): see for example Lacour (2008, Lemma 5, p. 812).

The following result of Baraud (2000) relies mainly on Talagrand's Inequality coupled with a moment inequality similar to Rosenthal's Inequality.

Proposition 4 (Baraud 2000, Corollary 5.1)**.** *Let $\widetilde{A}$ be a non-negative and symmetric matrix with at least a non-zero coefficient and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ a random vector in $\mathbb{R}^n$ with i.i.d. centered components. Assume that $\sigma^2 = \mathbb{E}[\varepsilon_1^2] < +\infty$ and set*

$$\zeta(\varepsilon) = \sqrt{\varepsilon'\widetilde{A}\varepsilon}.$$

For all $p \geq 2$ such that $\mathbb{E}[|\varepsilon_1|^p] < \infty$, we have, for all $x > 0$,

$$\mathbb{P}\left(\zeta^2(\varepsilon) \geq tr(\widetilde{A})\sigma^2 + 2\sigma^2\sqrt{\rho(\widetilde{A})tr(\widetilde{A})x} + \sigma^2\rho(\widetilde{A})x\right) \leq C\tau_p\frac{tr(\widetilde{A})}{\rho(\widetilde{A})x^{p/2}},$$

where $C > 0$ depends only on p and $\tau_p := \mathbb{E}[|\varepsilon_1|^p]/\sigma^p$.

Bibliographie

Aguilera, A. M., Escabias, M. et Valderrama, M. J. (2006). « Using principal components for estimating logistic regression with high-dimensional multicollinear data ». In : *Comput. Statist. Data Anal.* 50.8, p. 1905–1924 (cf. p. 24).

Aguilera, A. M., Escabias, M. et Valderrama, M. J. (2008). « Discussion of different logistic models with functional data. Application to systemic Lupus Erythematosus ». In : *Comput. Statist. Data Anal.* 53.1, p. 151–163 (cf. p. 24).

Aguilera, A., Aguilera-Morillo, M. C., Escabias, M. et Valderrama, M. (2011). « Penalized spline approaches for functional principal component logit regression ». In : *Recent advances in functional data analysis and related topics*. Contrib. Statist. Physica-Verlag/Springer, Heidelberg, p. 1–7 (cf. p. 24).

Aguilera-Morillo, M. C., Aguilera, A. M., Escabias, M. et Valderrama, M. J. (2013). « Penalized spline approaches for functional logit regression ». In : *TEST* 22.2, p. 251–277 (cf. p. 24).

Ait-Saïdi, A., Ferraty, F., Kassa, R. et Vieu, P. (2008). « Cross-validated estimations in the single-functional index model ». In : *Statistics* 42.6, p. 475–494 (cf. p. 24).

Allen, D. M. (1974). « The relationship between variable selection and data augmentation and a method for prediction ». In : *Technometrics* 16, p. 125–127 (cf. p. 8).

Alquier, P. et Wintenberger, O. (2012). « Model selection for weakly dependent time series forecasting ». In : *Bernoulli* 18.3, p. 883–913 (cf. p. 11).

Anderson, T. W. (1955). « The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities ». In : *Proc. Amer. Math. Soc.* 6, p. 170–176 (cf. p. 79).

Arlot, S. (2008). « V-fold cross-validation improved: V-fold penalization ». URL : `http://hal.archives-ouvertes.fr/hal-00239182` (cf. p. 8).

Arlot, S. et Celisse, A. (2010). « A survey of cross-validation procedures for model selection ». In : *Stat. Surv.* 4, p. 40–79 (cf. p. 8).

Arlot, S. et Lerasle, M. (2012). « V-fold cross-validation and V-fold penalization in least-squares density estimation ». URL : `http://hal.archives-ouvertes.fr/hal-00743931` (cf. p. 8).

Arlot, S. et Massart, P. (2009). « Data-driven Calibration of Penalties for Least-Squares Regression ». In : *J. Mach. Learn. Res.* 10, p. 245–279 (cf. p. 44).

Ash, R. B. et Gardner, M. F. (1975). *Topics in stochastic processes*. Probability and Mathematical Statistics, Vol. 27. New York : Academic Press [Harcourt Brace Jovanovich Publishers] (cf. p. 7, 36, 38, 39, 83, 86).

Aspirot, L., Bertin, K. et Perera, G. (2009). « Asymptotic normality of the Nadaraya-Watson estimator for nonstationary functional data and applications to telecommunications ». In : *J. Nonparametr. Stat.* 21.5, p. 535–551 (cf. p. 76).

Azzedine, N., Laksaci, A. et Ould-Saïd, E. (2008). « On robust nonparametric regression estimation for a functional regressor ». In : *Statist. Probab. Lett.* 78.18, p. 3216–3221 (cf. p. 76).

Baraud, Y. (2000). « Model selection for regression on a fixed design ». In : *Probab. Theory Related Fields* 117.4, p. 467–493 (cf. p. 58, 61, 181).

Baraud, Y. (2002). « Model selection for regression on a random design ». In : *ESAIM Probab. Statist.* 6, 127–146 (electronic) (cf. p. 9, 11, 17).

Baraud, Y., Giraud, C. et Huet, S. (2009). « Gaussian model selection with an unknown variance ». In : *Ann. Statist.* 37.2, p. 630–672 (cf. p. 17, 35).

Barron, A. R. (1994). « Approximation and estimation bounds for artificial neural networks ». In : *Machine Learning* 14.1, p. 115–133 (cf. p. 156).

Barron, A., Birgé, L. et Massart, P. (1999). « Risk bounds for model selection via penalization ». In : *Probab. Theory Related Fields* 113, p. 301–413 (cf. p. 9).

Bates, R. A., Buck, R. J., Riccomagno, E. et Wynn, H. P. (1996). « Experimental design and observation for large systems ». In : *J. Roy. Statist. Soc. Ser. B* 58.1. With discussion and a reply by the authors, p. 77–94, 95–111 (cf. p. 26).

Baudry, J.-P., Maugis, C. et Michel, B. (2012). « Slope heuristics: overview and implementation ». In : *Stat. Comput.* 22.2, p. 455–470 (cf. p. 10, 44).

Bayle, S., Monestiez, P. et Nerini, D. (2014). « Modèle linéaire de prédiction fonctionnelle sur données environnementales : choix de modélisation ». In : *J-SFdS* 155.2, p. 100–120 (cf. p. 31).

Benhenni, K., Ferraty, F., Rachdi, M. et Vieu, P. (2007). « Local smoothing regression with functional data ». In : *Comput. Statist.* 22.3, p. 353–369 (cf. p. 76).

Benhenni, K., Hedli-Griche, S. et Rachdi, M. (2010). « Estimation of the regression operator from functional fixed-design with correlated errors ». In : *J. Multivariate Anal.* 101.2, p. 476–490 (cf. p. 20).

Bertin, K., Lacour, C. et Rivoirard, V. (2013). *Adaptive estimation of conditional density function*. Rapp. tech. URL : `http://hal.archives-ouvertes.fr/hal-00922555` (cf. p. 13).

Birgé, L. et Massart, P. (1998). « Minimum contrast estimators on sieves: exponential bounds and rates of convergence ». In : *Bernoulli* 4.3, p. 329–375. ISSN : 1350-7265 (cf. p. 156, 180).

Birgé, L. et Massart, P. (2000). « An adaptive compression algorithm in Besov spaces ». In : *Constr. Approx.* 16.1, p. 1–36 (cf. p. 10).

Birgé, L. et Massart, P. (2007). « Minimal penalties for Gaussian model selection ». In : *Probab. Theory Related Fields* 138.1-2, p. 33–73 (cf. p. 10, 42, 44).

Bosq, D. (2000). *Linear processes in function spaces: theory and applications*. en. Springer (cf. p. 5, 159, 180).

Bouaziz, O. (2010). « Utilisation de modèles à direction révélatrice unique pour les modèles de durée ». Bosq, D. and Delecroix, M. (Dir.) Thèse de doct. Université Pierre et Marie Curie – Paris VI (cf. p. 24).

Box, G. E. P. et Hunter, J. S. (1957). « Multi-factor experimental designs for exploring response surfaces ». In : *Ann. Math. Stat.* 28.1, p. 195–241 (cf. p. 134).

Box, G. E. P. et Wilson, K. B. (1951). « On the experimental attainment of optimum conditions ». In : *J. Roy. Statist. Soc. Ser. B.* 13, p. 1–38, 1–38 (cf. p. 25, 128).

Brault, V., Baudry, J.-P., Maugis, C. et Michel, B. (2012). *capushe: Capushe, data-driven slope estimation and dimension jump*. R package version 1.0. URL : `http://CRAN.R-project.org/package=capushe` (cf. p. 44).

Brezis, H. (2005). *Analyse Fonctionnelle*. Sciences Sup. Dunod, Paris (cf. p. 3, 5, 158, 159).

Brook, R. J. et Arnold, G. C. (1985). *Applied Regression Analysis and Experimental Design*. CRC Press (cf. p. 132).

Brunel, E., Comte, F. et Guilloux, A. (2013). « Nonparametric estimation for survival data with censoring indicators missing at random ». In : *J. Statist. Plann. Inference* 143.10, p. 1653–1671 (cf. p. 11).

Brunel, E., Comte, F. et Lacour, C. (2010). « Minimax estimation of the conditional cumulative distribution function ». In : *Sankhya Ser. A* 72.2, p. 293–330 (cf. p. 79, 82, 85).

Burba, F., Ferraty, F. et Vieu, P. (2009). « k-Nearest Neighbour method in functional nonparametric regression ». In : *J. Nonparametr. Stat.* 21.4, p. 453–469 (cf. p. 79).

Cai, T. T. et Hall, P. (2006). « Prediction in Functional Linear Regression ». In : *Ann. Statist.* P. 2159–2179 (cf. p. 16, 31, 36, 37, 78).

Cai, T. T. et Yuan, M. (2012). « Minimax and adaptive prediction for functional linear regression ». In : *J. Amer. Statist. Assoc.* 107.499, p. 1201–1216 (cf. p. 16, 32, 37, 38).

Cardot, H., Cénac, P. et Zitt, P.-A. (2012). « Recursive estimation of the conditional geometric median in Hilbert spaces ». In : *Electron. J. Stat.* 6, p. 2535–2562 (cf. p. 3).

Cardot, H., Crambes, C. et Sarda, P. (2007). « Ozone pollution forecasting using conditional mean and conditional quantiles with functional covariates ». In : *Statistical methods for biostatistics and related fields*. Springer, Berlin, p. 221–243 (cf. p. 31).

Cardot, H., Ferraty, F. et Sarda, P. (1999). « Functional linear model ». In : *Stat. Probabil. Lett.* 45.1, p. 11–22 (cf. p. 13, 16, 40, 78).

Cardot, H., Ferraty, F. et Sarda, P. (2003). « Spline estimator for the Functional Linear Model ». In : *Stat. Sinica*, p. 571–591 (cf. p. 15, 16, 24, 48, 156).

Cardot, H. et Johannes, J. (2010). « Thresholding Projection Estimators in Functional Linear Models. » In : *J. Multivariate Anal.* P. 395–408 (cf. p. 15, 16, 19, 31, 32, 37, 38).

Cardot, H., Mas, A. et Sarda, P. (2007). « CLT in functional linear regression models ». In : *Probab. Theory Related Fields* 138.3-4, p. 325–361 (cf. p. 6, 158, 170).

Cardot, H. et Sarda, P. (2005). « Estimation in generalized linear models for functional data via penalized likelihood ». In : *J. Multivariate Anal.* 92.1, p. 24–41 (cf. p. 24).

Cavalier, L. (2011). « Inverse problems in statistics ». In : *Inverse problems and high-dimensional estimation*. T. 203. Lect. Notes Stat. Proc. Springer, Heidelberg, p. 3–96 (cf. p. 14).

Cavalier, L. et Hengartner, N. W. (2005). « Adaptive estimation for inverse problems with noisy operators ». In : *Inverse Problems* 21.4, p. 1345–1361 (cf. p. 15).

Celisse, A. (2008). « Optimal cross-validation in density estimation ». URL : `http://hal.archives-ouvertes.fr/hal-00337058` (cf. p. 8).

Chagny, G. (2013a). « Estimation adaptative avec des données transformées ou incomplètes. Application à des modèles de survie ». Thèse de doct. Univ. Paris Descartes (cf. p. 81).

Chagny, G. (2013b). « Penalization versus Goldenshluger-Lepski strategies in warped bases regression ». In : *ESAIM Probab. Statist.* 17, 328–358 (electronic) (cf. p. 13, 22, 81).

Charlier, B. (2011). « Étude des propriétés statistiques des moyennes de Fréchet dans des modèles de déformations pour l'analyse de courbes et d'images en grande dimension ». Thèse de doct. Univ. Toulouse (cf. p. 157).

Chen, D., Hall, P. et Müller, H.-G. (2011). « Single and multiple index functional regression models with nonparametric link ». In : *Ann. Statist.* 39.3, p. 1720–1747 (cf. p. 24).

Cho, H., Goude, Y., Brossat, X. et Yao, Q. (2013). « Modeling and Forecasting Daily Electricity Load Curves: A Hybrid Approach ». In : *J. Amer. Statist. Assoc.* 108.501, p. 7–21 (cf. p. 31).

Comte, F. et Genon-Catalot, V. (2012). « Convolution power kernels for density estimation ». In : *J. Statist. Plann. Inference* 142.7, p. 1698–1715 (cf. p. 99).

Comte, F. et Johannes, J. (2010). « Adaptive Estimation in Circular Functional Linear Models ». In : *Math. Method Statist.* P. 42–63 (cf. p. 6, 15, 16, 31, 32, 37, 152).

Comte, F. et Johannes, J. (2012). « Adaptive functional linear regression ». In : *Ann. Statist.* 40.6, p. 2765–2797 (cf. p. 13, 22, 31, 32, 81).

Comte, F. et Lacour, C. (2010). « Pointwise deconvolution with unknown error distribution ». In : *C. R. Math. Acad. Sci. Paris* 348.5-6, p. 323–326 (cf. p. 85).

Comte, F. et Lacour, C. (2013). « Anisotropic adaptive kernel deconvolution ». In : *Annales de l'IHP* 49.2, p. 569–609 (cf. p. 13).

Crambes, C., Delsol, L. et Laksaci, A. (2008). « Robust nonparametric estimation for functional data ». In : *J. Nonparametr. Stat.* 20.7, p. 573–598 (cf. p. 20, 76).

Crambes, C., Kneip, A. et Sarda, P. (2009). « Smoothing Splines Estimators for Functional Linear Regression ». In : *Ann. Statist.* P. 35–72 (cf. p. 15, 16, 31, 32, 37, 52, 54, 78).

Crambes, C. et Mas, A. (2012). « Asymptotics of prediction in functional linear regression with functional outputs ». Anglais. In : *Bernoulli* 19.5B, p. 2627–2651 (cf. p. 158, 170).

Dabo-Niang, S. (2004). « Kernel density estimator in an infinite-dimensional space with a rate of convergence in the case of diffusion process ». In : *Appl. Math. Lett.* 17.4, p. 381–386 (cf. p. 8).

Dabo-Niang, S., Kaid, Z. et Laksaci, A. (2012). « On spatial conditional mode estimation for a functional regressor ». In : *Statist. Probab. Lett.* 82.7, p. 1413–1421 (cf. p. 76).

Dabo-Niang, S. et Rhomari, N. (2009). « Kernel regression estimation in a Banach space ». In : *J. Statist. Plann. Inference* 139.4, p. 1421–1434 (cf. p. 76).

Dabo-Niang, S. et Yao, A.-F. (2013). « Kernel spatial density estimation in infinite dimension space ». In : *Metrika* 76.1, p. 19–52 (cf. p. 8, 78).

Dauxois, J., Pousse, A. et Romain, Y. (1982). « Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference ». In : *J. Multivariate Anal.* 12.1, p. 136–154 (cf. p. 5, 6).

Deheuvels, P. (2007). « A Karhunen-Loève expansion for a mean-centered Brownian bridge ». In : *Statist. Probab. Lett.* 77.12, p. 1190–1200 (cf. p. 7).

Delaigle, A. et Hall, P. (2010). « Defining probability density for a distribution of random functions ». In : *Ann. Statist.* 38.2, p. 1171–1193 (cf. p. 7, 78).

Delaigle, A. et Hall, P. (2012). « Methodology and theory for partial least squares applied to functional data ». In : *Ann. Statist.* 40.1, p. 322–352 (cf. p. 138, 139).

Delecroix, M. et Hristache, M. (1999). « M-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique ». In : *Bull. Belg. Math. Soc. Simon Stevin* 6.2, p. 161–185 (cf. p. 24).

Delsol, L. (2010). « Régression sur variable fonctionnelle: Estimation, tests de structure et Applications ». Thèse de doct. Univ. Toulouse III (cf. p. 20, 23).

DeVore, R. A. et Lorentz, G. G. (1993). *Constructive Approximation*. Grundelehren der matematischen Wissenschaften 303. Springer-Verlag Berlin Heidelberg (cf. p. 10, 152).

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S. et Punjabi, N. M. (2009). « Multilevel functional principal component analysis ». In : *Ann. Appl. Stat.* 3.1, p. 458–488 (cf. p. 3).

Doumic, M., Hoffmann, M., Reynaud-Bouret, P. et Rivoirard, V. (2012). « Nonparametric estimation of the division rate of a size-structured population ». In : *SIAM J. Numer. Anal.* 50.2, p. 925–950 (cf. p. 13).

Draper, N. R. et Guttman, I. (1988). « An index of rotatability ». In : *Technometrics* 30.1, p. 105–111 (cf. p. 135).

Draper, N. R. et Pukelsheim, F. (1990). « Another look at rotatability ». In : *Technometrics* 32.2, p. 195–202 (cf. p. 135).

Dunford, N. et Schwartz, J. T. (1958). *Linear Operators. I. General Theory*. With the assistance of W. G. Bade and R. G. Bartle. Pure and Applied Mathematics, Vol. 7. Interscience Publishers, Inc., New York (cf. p. 4, 161).

Dunker, T., Lifshits, M. A. et Linde, W. (1998). « Small deviation probabilities of sums of independent random variables ». In : *High dimensional probability (Oberwolfach, 1996)*. T. 43. Progr. Probab. Basel : Birkhäuser, p. 59–74 (cf. p. 83).

Escabias, M., Aguilera, A. M. et Valderrama, M. J. (2004). « Principal component estimation of functional logistic regression: discussion of two different approaches ». In : *J. Nonparametr. Stat.* 16.3-4, p. 365–384 (cf. p. 24).

Escabias, M., Aguilera, A. M. et Valderrama, M. J. (2005). « Modeling environmental data by functional principal component logistic regression ». In : *Environmetrics* 16.1, p. 95–107 (cf. p. 24).

Facer, M. R. et Müller, H.-G. (2003). « Nonparametric estimation of the location of a maximum in a response surface ». In : *J. Multivariate Anal.* 87.1, p. 191–217 (cf. p. 26).

Ferraty, F., Laksaci, A., Tadj, A. et Vieu, P. (2010). « Rate of uniform consistency for nonparametric estimates with functional variables ». In : *J. Stat. Plan. Infer.* 2, p. 335–352 (cf. p. 20, 76, 79, 80, 87, 101, 153).

Ferraty, F., Laksaci, A. et Vieu, P. (2006). « Estimating some characteristics of the conditional distribution in nonparametric functional models ». In : *Stat. Infer. Stoch. Proc.* 9.1 (cf. p. 20–22, 74, 76, 77, 79, 80, 84, 87, 101, 153).

Ferraty, F., Mas, A. et Vieu, P. (2007). « Nonparametric Regression on Functional Data: Inference and Practical Aspects ». In : *Aust. NZ J. Stat.* 49.3, p. 267–286 (cf. p. 76, 97).

Ferraty, F. et Romain, Y. (2011). *The Oxford Handbook of Functional Data Analysis*. Oxford Handbooks in Mathematics. OUP Oxford (cf. p. 3).

Ferraty, F. et Vieu, P. (2002). « The functional nonparametric model and application to spectrometric data ». In : *Comput. Statist.* 17.4 (cf. p. 19, 76, 97).

Ferraty, F. et Vieu, P. (2004). « Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination ». In : *J. Nonparametr. Stat.* 16.1-2. The International Conference on Recent Trends and Directions in Nonparametric Statistics, p. 111–125 (cf. p. 20).

Ferraty, F. et Vieu, P. (2006). *Nonparametric functional data analysis*. Springer Series in Statistics. Theory and practice. New York : Springer (cf. p. 3, 79, 97).

Fischer, A. (2013). « Selecting the length of a principal curve within a Gaussian model ». In : *Electron. J. Stat.* 7, p. 342–363 (cf. p. 11).

Friedman, J. H. et Stuetzle, W. (1981). « Projection pursuit regression ». In : *J. Amer. Statist. Assoc.* 76.376, p. 817–823 (cf. p. 24).

Geenens, G. (2011). « Curse of dimensionality and related issues in nonparametric functional regression ». In : *Stat. Surv.* 5, p. 30–43 (cf. p. 20, 23, 77, 153, 157).

Geisser, S. (1975). « The Predictive Sample Reuse Method with Applications ». In : *J. Amer. Statist. Assoc.* 70.350, p. 320–328 (cf. p. 8).

Georgiou, S. D., Stylianou, S. et Aggarwal, M. (2014). « A class of composite designs for response surface methodology ». In : *Comput. Statist. Data Anal.* 71, p. 1124–1133 (cf. p. 26, 131).

Gheriballah, A., Laksaci, A. et Sekkal, S. (2013). « Nonparametric M-regression for functional ergodic data ». In : *Statist. Probab. Lett.* 83.3, p. 902–908 (cf. p. 76).

Giraud, C., Huet, S. et Verzelen, N. (2012). « Graph selection with GGMselect ». In : *Stat. Appl. Genet. Mol. Biol.* 11.3, p. 1544–6115 (cf. p. 11).

Goldenshluger, A. et Lepski, O. (2011). « Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality ». In : *Ann. Statist.* 39.3, p. 1608–1632 (cf. p. 9, 11, 12, 16, 21, 22, 32, 74, 76, 81, 82, 157).

Goldenshluger, A. et Lepski, O. (2013a). « General selection rule from a family of linear estimators ». In : *Theory Probab. Appl.* 57.2, p. 209–226 (cf. p. 13).

Goldenshluger, A. et Lepski, O. (2013b). « On adaptive minimax density estimation on R^d ». In : *Probab. Theory Related Fields* to appear. (Cf. p. 81).

Gunst, R. F. et Mason, R. L. (2009). « Fractional factorial design ». In : *WIREs Comp. Stat.* 1.2, p. 234–244 (cf. p. 131).

Hadamard, J. (1902). « Sur les problèmes aux dérivées partielles et leur signification physique ». In : *Princeton University Bulletin*, p. 49–52 (cf. p. 14).

Hall, P. (2011). « Principal component analysis for functional data: methodology, theory, and discussion ». In : *The Oxford handbook of functional data analysis*. Oxford : Oxford Univ. Press, p. 210–234 (cf. p. 5).

Hall, P. et Horowitz, J. L. (2007). « Methodology and Convergence Rates for Functional Linear Regression ». In : *Ann. Statist.* P. 70–91 (cf. p. 16, 36).

Hall, P. et Hosseini-Nasab, M. (2006). « On properties of functional principal components analysis ». In : *J. Roy. Stat. B*, p. 109–126 (cf. p. 6, 16, 36, 39, 48).

Halmos, P. R. (1963). « What does the spectral theorem say? » In : *Amer. Math. Monthly*, p. 241–247 (cf. p. 159).

Hastie, T., Tibshirani, R. et Friedman, J. (2009). *The elements of statistical learning*. Second. Springer Series in Statistics. Data mining, inference, and prediction. New York : Springer (cf. p. 8).

Hilgert, N., Mas, A. et Verzelen, N. (2013). « Minimax adaptive tests for the functional linear model ». In : *Ann. Statist.* 41.2, p. 838–869 (cf. p. 6, 165).

Hoffmann-Jørgensen, J., Shepp, L. A. et Dudley, R. M. (1979). « On the lower tail of Gaussian seminorms ». In : *Ann. Probab.* 7.2, p. 319–342 (cf. p. 79, 83).

Ibragimov, I. A. et Rozanov, Y. A. (1978). *Gaussian random processes*. T. 9. Applications of Mathematics. Translated from the Russian by A. B. Aries. Springer-Verlag, New York-Berlin, p. x+275 (cf. p. 6).

Ichimura, H. (1993). « Semiparametric least squares (SLS) and weighted SLS estimation of single-index models ». In : *J. Econometrics* 58.1-2, p. 71–120 (cf. p. 24).

James, G. M. (2002). « Generalized linear models with functional predictors ». In : *J. R. Stat. Soc. Ser. B Stat. Methodol.* 64.3, p. 411–432 (cf. p. 23).

James, G. M. et Silverman, B. W. (2005). « Functional Adaptive Model Estimation ». In : 100.470, p. 565–576 (cf. p. 24).

Kato, T. (1995). *Perturbation theory for linear operators*. Classics in Mathematics. Reprint of the 1980 edition. Berlin : Springer-Verlag (cf. p. 158).

Kerkyacharian, G., Lepski, O. et Picard, D. (2001). « Nonlinear estimation in anisotropic multi-index denoising ». In : *Probab. Theory Related Fields* 121.2, p. 137–170 (cf. p. 81).

Khuri, A. I. (1988). « A measure of rotatability for response-surface designs ». In : *Technometrics* 30.1, p. 95–104 (cf. p. 135).

Khuri, A. I. (2001). « An overview of the use of generalized linear models in response surface methodology ». In : *Proceedings of the Third World Congress of Nonlinear Analysts, Part 3 (Catania, 2000)*. T. 47. 3, p. 2023–2034 (cf. p. 26).

Khuri, A. I. et Mukhopadhyay, S. (2010). « Response surface methodology ». In : *Wiley Interdiscip. Rev. Comput. Stat.* 2.2, p. 128–149 (cf. p. 26).

Klein, T. et Rio, E. (2005). « Concentration around the mean for maxima of empirical processes ». In : *Ann. Probab.* 33.3, p. 1060–1077 (cf. p. 181).

Koltchinskii, V. et Giné, E. (2000). « Random matrix approximation of spectra of integral operators ». In : *Bernoulli* 6.1, p. 113–167 (cf. p. 158).

Lacour, C. (2006). « Rates of convergence for nonparametric deconvolution ». In : *C. R. Math. Acad. Sci. Paris* 342.11, p. 877–882 (cf. p. 85).

Lacour, C. (2008). « Adaptive estimation of the transition density of a particular hidden Markov chain ». In : *J. Multivariate Anal.* 99.5, p. 787–814 (cf. p. 181).

Laib, N. et Louani, D. (2010). « Nonparametric kernel regression estimation for functional stationary ergodic data: asymptotic properties ». In : *J. Multivariate Anal.* 101.10, p. 2266–2281 (cf. p. 76).

Laurini, M. P. (2014). « Dynamic functional data analysis with non-parametric state space models ». In : *J. Appl. Stat.* 41.1, p. 142–163 (cf. p. 3).

Lee, J. et Hajela, P. (1996). « Parallel genetic algorithm implementation in multidisciplinary rotor blade design ». In : *J. Aircraft* 33.5, p. 962–969 (cf. p. 26).

Lee, S.-Y., Zhang, W. et Song, X.-Y. (2002). « Estimating the covariance function with functional data ». In : *Br. J. Math. Stat. Psych.* 55.2, p. 247–261 (cf. p. 6).

Lenth, R. V. (2009). « Response-Surface Methods in R, Using rsm ». In : *J. Statist. Software* 32.7, p. 1–17 (cf. p. 141).

Lepski, O. et Serdyukova, N. (2014). « Adaptive estimation under single-index constraint in a regression model ». In : *Ann. Statist.* 42.1, p. 1–28 (cf. p. 13).

Lerasle, M. (2012). « Optimal model selection in density estimation ». In : *Ann. Inst. Henri Poincaré Probab. Stat.* 48.3, p. 884–908 (cf. p. 44).

Li, W. V. et Shao, Q.-M. (2001). « Gaussian processes: inequalities, small ball probabilities and applications ». In : *Stochastic processes: theory and methods*. T. 19. Handbook of Statist. Amsterdam : North-Holland, p. 533–597 (cf. p. 20, 79).

Li, Y. et Hsing, T. (2007). « On rates of convergence in functional linear regression ». In : *J. Multivariate Anal.* 98.9, p. 1782–1804 (cf. p. 15, 52).

Li, Y., Wang, N. et Carroll, R. J. (2010). « Generalized functional linear models with semiparametric single-index interactions ». In : *J. Amer. Statist. Assoc.* 105.490, p. 621–633 (cf. p. 24).

Lian, H. (2012). « Convergence of nonparametric functional regression estimates with functional responses ». In : *Electron. J. Stat.* 6, p. 1373–1391 (cf. p. 20).

MacNeill, I. B. (1978). « Properties of sequences of partial sums of polynomial regression residuals with applications to tests for change of regression at unknown times ». In : *Ann. Statist.* 6.2, p. 422–433 (cf. p. 7).

Mallows, C. L. (1973). « Some Comments on C_P ». In : *Technometrics* 15.4, p. 661–675 (cf. p. 10).

Marx, B. D. et Eilers, P. H. (1996). « Flexible Smoothing with B-splines and Penalties ». In : *Statist. Sci.* P. 89–121 (cf. p. 48).

Mas, A. (2012). « Lower bound in regression for functional data by representation of small ball probabilities ». In : *Electron. J. Statist.* 6, p. 1745–1778 (cf. p. 20, 78, 83–85, 87, 154).

Mas, A. et Ruymgaart, F. (2014). « High-Dimensional Principal Projections ». In : *Complex Anal. Oper. Theory.* to appear (cf. p. 6, 165).

Masry, E. (2005). « Nonparametric regression estimation for dependent functional data: asymptotic normality ». In : *Stochastic Process. Appl.* 115.1, p. 155–177 (cf. p. 76, 77).

Massart, P. (2007). *Concentration inequalities and model selection.* T. 1896. Lecture Notes in Mathematics. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. Berlin : Springer (cf. p. 9, 11, 17).

Montgomery, D. C. (2009). *Design and analysis of experiments.* Seventh. Hoboken, NJ : John Wiley & Sons Inc. (cf. p. 25, 130).

Mörters, P. et Peres, Y. (2010). *Brownian motion.* T. 30. Cambridge Series in Statistical and Probabilistic Mathematics. With an appendix by Oded Schramm and Wendelin Werner. Cambridge University Press, Cambridge, p. xii+403. ISBN : 978-0-521-76018-8 (cf. p. 8).

Müller, H.-G. et Stadtmüller, U. (2005). « Generalized functional linear models ». In : *Ann. Statist.* 33.2, p. 774–805 (cf. p. 23).

Myers, R. H., Montgomery, D. C. et Anderson-Cook, C. M. (2009). *Response surface methodology.* Third. Wiley Series in Probability and Statistics. Process and product optimization using designed experiments. Hoboken, NJ : John Wiley & Sons Inc. (cf. p. 132, 134).

Nemirovski, A. (2000). « Topics in non-parametric statistics ». In : *Lectures on probability theory and statistics (Saint-Flour, 1998).* T. 1738. Lecture Notes in Math. Berlin : Springer, p. 85–277 (cf. p. 8).

Newey, W. K. et Stoker, T. M. (1993). « Efficiency of Weighted Average Derivative Estimators and Index Models ». In : *Econometrica* 61.5, p. 1199–1223 (cf. p. 24).

Park, S. H., Lim, J. H. et Baba, Y. (1993). « A measure of rotatability for second order response surface designs ». In : *Ann. Inst. Statist. Math.* 45.4, p. 655–664 (cf. p. 135).

Pázman, A. (1986). *Foundations of optimum experimental design.* T. 14. Mathematics and its Applications (East European Series). Translated from the Czech. Dordrecht : D. Reidel Publishing Co. (cf. p. 135).

Plancade, S. (2013). « Adaptive estimation of the conditional cumulative distribution function from current status data ». In : *J. Statist. Plann. Inference* 143.9, p. 1466–1485 (cf. p. 85).

Preda, C. et Saporta, G. (2005). « PLS regression on a stochastic process ». In : *Comput. Statist. Data Anal.* 48.1, p. 149–158 (cf. p. 138, 156).

Rachdi, M. et Vieu, P. (2007). « Nonparametric regression for functional data: automatic smoothing parameter selection ». In : *J. Statist. Plann. Inference* 137.9, p. 2784–2801 (cf. p. 76).

Rakêt, L. L. et Markussen, B. (2014). « Approximate inference for spatial functional data on massively parallel processors ». In : *Comput. Statist. Data Anal.* 72, p. 227–240 (cf. p. 3).

Ramsay, J. O. et Dalzell, C. J. (1991). « Some Tools for Functional Data Analysis ». In : *J. Roy. Stat. Soc. B Met.* P. 539–572 (cf. p. 13, 15).

Ramsay, J. et Silverman, B. (2005). *Functional Data Analysis*. anglais. 2^{e} éd. Springer Series in Statistics. Springer (cf. p. 3, 5, 6, 15, 16, 32, 40, 41, 78, 152, 156).

Reiss, P. T. et Ogden, R. T. (2007). « Functional principal component regression and functional partial least squares ». In : *J. Amer. Statist. Assoc.* 102.479, p. 984–996 (cf. p. 15).

Rice, J. A. et Silverman, B. W. (1991). « Estimating the mean and covariance structure nonparametrically when the data are curves ». In : *J. Roy. Statist. Soc. Ser. B* 53.1, p. 233–243 (cf. p. 6).

Rudin, W. (1973). *Functional analysis*. McGraw-Hill Series in Higher Mathematics. New York : McGraw-Hill Book Co., p. xiii+397 (cf. p. 158, 160).

Rudin, W. (1987). *Real and complex analysis*. Third. New York : McGraw-Hill Book Co., p. xiv+416. ISBN : 0-07-054234-1 (cf. p. 161, 162).

Sacks, J., Welch, W. J., Mitchell, T. J. et Wynn, H. P. (1989). « Design and analysis of computer experiments ». In : *Statist. Sci.* 4.4. With comments and a rejoinder by the authors, p. 409–435 (cf. p. 26).

Sauder, C., Cardot, H., Disenhaus, C. et Le Cozler, Y. (2013). « Non-parametric approaches to the impact of Holstein heifer growth from birth to insemination on their dairy performance at lactation one ». In : *J. Agr. Sci.* 151, p. 578–589 (cf. p. 3).

Schwarz, G. (1978). « Estimating the dimension of a model ». In : *Ann. Statist.* 6.2, p. 461–464 (cf. p. 8).

Shang, H. L. (2013). « Bayesian bandwidth estimation for a nonparametric functional regression model with unknown error density ». In : *Comput. Statist. Data Anal.* 67, p. 185–198 (cf. p. 76).

Simon, B. (2005). *Trace, ideals and their applications*. Second. T. 120. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, p. viii+150 (cf. p. 5).

Sørensen, H., Tolver, A., Thomsen, M. H. et Andersen, P. H. (2012). « Quantification of symmetry for functional data with application to equine lameness classification ». In : *J. Appl. Statist.* 39.2, p. 337–360 (cf. p. 3).

Stone, M. (1974). « Cross-validatory choice and assessment of statistical predictions ». In : *J. Roy. Statist. Soc. Ser. B* 36. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors, p. 111–147 (cf. p. 8).

Tibshirani, R. (1996). « Regression shrinkage and selection via the lasso ». In : *J. Roy. Statist. Soc. Ser. B* 58.1, p. 267–288 (cf. p. 8).

Tsybakov, A. B. (2008). « Agrégation d'estimateurs et optimisation stochastique ». In : *J. Soc. Fr. Stat. & Rev. Stat. Appl.* 149.1, p. 3–26 (cf. p. 8).

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Revised and extended from the 2004 French original, Translated by V. Zaiats. New York : Springer, p. xii+214 (cf. p. 12, 37, 115, 116, 120, 122).

Verzelen, N. (2010). « Data-driven neighborhood selection of a Gaussian field ». In : *Comput. Statist. Data Anal.* 54.5, p. 1355–1371 (cf. p. 44).

Wold, H. (1975). « Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach ». In : *Perspectives in probability and statistics (papers in honour of M. S. Bartlett on the occasion of his 65th birthday)*. Univ. Sheffield, Sheffield : Applied Probability Trust, p. 117–142 (cf. p. 138).

Yang, Y. (2003). « Regression with multiple candidate models: selecting or mixing? » In : *Statist. Sinica* 13.3, p. 783–809 (cf. p. 8).

Zou, H. et Hastie, T. (2005). « Regularization and variable selection via the elastic net ». In : *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67.2, p. 301–320 (cf. p. 8).

Zwald, L. et Blanchard, G. (2005). « On the Convergence of Eigenspaces in Kernel Principal Component Analysis ». In : *Advances in Neural Information Processing Systems 18, NIPS 2005*. Vancouver, British Columbia, Canada, p. 1–8 (cf. p. 158).

Modélisation statistique pour données fonctionnelles : approches non-asymptotiques et méthodes adaptatives

Résumé : L'objet principal de cette thèse est de développer des estimateurs adaptatifs en statistique pour données fonctionnelles. Dans une première partie, nous nous intéressons au modèle linéaire fonctionnel et nous définissons un critère de sélection de la dimension pour des estimateurs par projection définis sur des bases fixe ou aléatoire. Les estimateurs obtenus vérifient une inégalité de type oracle et atteignent la vitesse de convergence minimax pour le risque lié à l'erreur de prédiction. Pour les estimateurs définis sur une collection de modèles aléatoires, des outils de théorie de la perturbation ont été utilisés pour contrôler les projecteurs aléatoires de manière non-asymptotique. D'un point de vue numérique, cette méthode de sélection de la dimension est plus rapide et plus stable que les méthodes usuelles de validation croisée. Dans une seconde partie, nous proposons un critère de sélection de fenêtre inspiré des travaux de Goldenshluger et Lepski, pour des estimateurs à noyau de la fonction de répartition conditionnelle lorsque la covariable est fonctionnelle. Le risque de l'estimateur obtenu est majoré de manière non-asymptotique. Des bornes inférieures sont prouvées ce qui nous permet d'établir que notre estimateur atteint la vitesse de convergence minimax, à une perte logarithmique près. Dans une dernière partie, nous proposons une extension au cadre fonctionnel de la méthodologie des surfaces de réponse, très utilisée dans l'industrie. Ce travail est motivé par une application à la sûreté nucléaire.

Mots clés : données fonctionnelles, estimateurs adaptatifs, régression, sélection de modèle, méthode de Goldenshluger-Lepski, méthodologie des surfaces de réponses.

Statistical modeling for functional data: non-asymptotic approaches and adaptive methods

Abstract: The main purpose of this thesis is to develop adaptive estimators for functional data. In the first part, we focus on the functional linear model and we propose a dimension selection device for projection estimators defined on both fixed and data-driven bases. The prediction error of the resulting estimators satisfies an oracle-type inequality and reaches the minimax rate of convergence. For the estimator defined on a data-driven approximation space, tools of perturbation theory are used to solve the problems related to the random nature of the collection of models. From a numerical point of view, this method of dimension selection is faster and more stable than the usual methods of cross validation. In a second part, we consider the problem of bandwidth selection for kernel estimators of the conditional cumulative distribution function when the covariate is functional. The method is inspired by the work of Goldenshluger and Lepski. The risk of the estimator is non-asymptotically upper-bounded. We also prove lower-bounds and establish that our estimator reaches the minimax convergence rate, up to an extra logarithmic term. In the last part, we propose an extension to a functional context of the response surface methodology, widely used in the industry. This work is motivated by an application to nuclear safety.

Keywords: functional data analysis, adaptive estimators, regression, model selection, Goldenshluger and Lepski's method, response surface methodology.